

**A peer-reviewed version of this preprint was published in PeerJ on 19 April 2016.**

[View the peer-reviewed version](https://doi.org/10.7717/peerj.1958) (peerj.com/articles/1958), which is the preferred citable publication unless you specifically need to cite this preprint.

Anuwongcharoen N, Shoombuatong W, Tantimongcolwat T, Prachayasittikul V, Nantasenamat C. 2016. Exploring the chemical space of influenza neuraminidase inhibitors. PeerJ 4:e1958  
<https://doi.org/10.7717/peerj.1958>

# Exploring the chemical space of influenza neuraminidase inhibitors

Nuttapat Anuwongcharoen, Watshara Shoombuatong, Tanawut Tantimongcolwat, Virapong Prachayasittikul, Chanin Nantasenamat

The combat against the emergence of mutant influenza strains has led to the screening of a growing number of compounds for inhibitory activity against influenza neuraminidase. This study explores the chemical space of neuraminidase inhibitors (NAIs) provides an opportunity for further gaining molecular insights on the underlying basis of the bioactivity. Particularly, a large set of 347 and 175 NAIs against influenza A and B, respectively, was compiled from the literature. Molecular and quantum chemical descriptors were obtained from low-energy conformational structures geometrically optimized at B3LYP/6-31G(d) level. The bioactivity of NAIs were classified as active and inactive NAIs according to their half maximum inhibitory concentration ( $\text{IC}_{50}$ ) value in which  $\text{IC}_{50} < 1 \mu\text{M}$  and  $> 10 \mu\text{M}$  were defined as active and inactive compounds against influenza neuraminidase, respectively. Interpretable decision rules were derived from a quantitative structure-activity relationship (QSAR) model established using 13 descriptors by means of decision tree analysis. Good predictive performance was achieved as deduced from ten-fold cross-validation where accuracy and MCC of 87.5% and 0.731, respectively, were obtained for influenza A NAIs while values of 89.78% and 0.786 for influenza B NAIs. Both univariate and multivariate analyses revealed the importance of lowest unoccupied molecular orbital, number of hydrogen bond donor and number of hydrogen bond acceptors in the predictive model of NAIs against influenza A while the number of hydrogen bond donor, number of hydrogen bond acceptor and energy gap between highest occupied and lowest unoccupied molecular orbital were important in the predictive model for influenza B NAIs. Analysis of molecular scaffold was performed on both data sets in combination with functional group analysis for discriminating important structural features amongst active and inactive NAIs. Furthermore, molecular docking was deployed to investigate the binding mode and their moiety preferences of active NAIs against both influenza A and B neuraminidase. Results from this study is anticipated to be beneficial for guiding the rational drug design of novel NAIs for treatment of influenza infection.

# Exploring the chemical space of influenza neuraminidase inhibitors

Nuttapat Anuwongcharoen<sup>1,2</sup>, Watshara Shoombuatong<sup>1</sup>,  
Tanawut Tantimongkolwat<sup>3</sup>, Virapong Prachayasittikul<sup>2</sup>, and  
Chanin Nantasenamat<sup>\*1</sup>

<sup>1</sup>Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology,  
Mahidol University, Bangkok 10700, Thailand

<sup>2</sup>Department of Clinical Microbiology and Applied Technology, Faculty of Medical  
Technology, Mahidol University, Bangkok 10700, Thailand

<sup>3</sup>Center for Research and Innovation, Faculty of Medical Technology, Mahidol  
University, Bangkok 10700, Thailand

## ABSTRACT

The fight against the emergence of mutant influenza strains has led to the screening of an increasing number of compounds for inhibitory activity against influenza neuraminidase. This study explores the chemical space of neuraminidase inhibitors (NAIs), which provides an opportunity to obtain further molecular insights regarding the underlying basis of the bioactivity. In particular, a large set of 347 and 175 NAIs against influenza A and B, respectively, was compiled from the literature. Molecular and quantum chemical descriptors were obtained from low-energy conformational structures geometrically optimized at the B3LYP/6-31G(d) level. The bioactivities of the NAIs were classified as active or inactive according to their half maximum inhibitory concentration (IC<sub>50</sub>) value, in which IC<sub>50</sub> < 1 μM and > 10 μM were defined as active and inactive compounds against influenza neuraminidase, respectively. Interpretable decision rules were derived from a quantitative structure-activity relationship (QSAR) model established using 13 descriptors via decision tree analysis. Good predictive performance was achieved, as deduced from ten-fold cross-validation, in which an accuracy and MCC of 87.5% and 0.731, respectively, were obtained for influenza A NAIs, while values of 89.78% and 0.786 were obtained for influenza B NAIs. Both univariate and multivariate analyses revealed the importance of the lowest unoccupied molecular orbital, number of hydrogen bond donors and number of hydrogen bond acceptors in the predictive model of NAIs against influenza A, while the number of hydrogen bond donors, number of hydrogen bond acceptors and the energy gap between the highest occupied and lowest unoccupied molecular orbitals were important in the predictive model for influenza B NAIs. Molecular scaffold analysis was performed on both data sets in combination with functional group analysis for discriminating important structural features among active and inactive NAIs. Furthermore, molecular docking was employed to investigate the binding modes and their moiety preferences of active NAIs against both influenza A and B neuraminidase. The results from this study are anticipated to be beneficial for guiding the rational drug design of novel NAIs for treating influenza infections.

**Keywords:** influenza, neuraminidase, neuraminidase inhibitor, chemical space, QSAR, scaffold analysis, fragment analysis, functional group analysis, molecular docking, rational drug design

## INTRODUCTION

Influenza is one of the most concerning diseases for global public health, and it is caused by influenza viruses, which are enveloped segmented-RNA viruses that belong to the Orthomyxoviridae family. The global estimate for cases of seasonal influenza infection is as high as 1 billion cases per year, in which approximately 3 to 5 million cases often develop a progressive severe illness and lead to 250,000 to 500,000 fatalities per year worldwide (World Health Organization, 2014). Among the severe cases, high fatality rates are observed particularly in very young children and elderly people >65 years of age, who are considered to be a risk group vulnerable to influenza infection. Thus, influenza infections significantly

\*Corresponding author. E-mail: chanin.nan@mahidol.ac.th

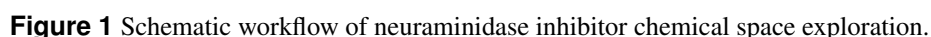
increase the number of hospitalizations, lead to substantial economical losses from disease intervention and impact the productivity of society (Peasah et al., 2013).

The current strategy for treating influenza focuses on inhibiting the function of neuraminidase, which is an enveloped enzyme located on the surface of both influenza type A and B. Influenza neuraminidase is an exosialidase that recognizes the  $\alpha$ -ketosidic linkage between neuraminic acid (or sialic acid) and carbohydrate residues (von Itzstein, 2011). The influenza virus requires this enzyme to facilitate viral budding of progeny virions out of the cells and to prevent viral aggregation of virus particles. The interaction allows the mature virus to detach from the host cell, resulting in the release of progeny virions from the surface of the host cell. Moreover, neuraminidase also plays a role in the cleavage of neuraminic acid of mucin inside the respiratory tract, thereby facilitating the movement of the virus toward its target cells (Shtyrya et al., 2009). Thus, neuraminidase is a crucial enzyme that facilitates viral spreading and transmission. To prevent the spreading of influenza viruses, neuraminidase inhibitors (NAIs) are currently an effective choice for treatment and prophylaxis.

Currently, only three NAIs have been approved for use as therapeutic and prophylaxis agents of influenza virus: zanamivir (Relenza), oseltamivir (Tamiflu) and peramivir (Rapivab). Zanamivir is the first approved nasally administered NAI, and it exerts highly effective inhibitory activity against both types of influenza virus. This dihydropyran-based NAI was developed based on the structural modification of a sialic acid analogue called DANA (Meindl et al., 1974). Due to its high polarity, zanamivir exhibits low oral bioavailability and requires administration via nasal inhalation. Oseltamivir is a second-generation NAI approved for use as an oral anti-influenza agent, and it exhibits efficacy comparable to that of zanamivir (Tuna et al., 2012). This cyclohexene-based NAI is less polar than the previous generation, thus making it easier to administer than the inhalation route. The most recently approved intravenous NAI, peramivir, was announced in December 2014. This intravenous formulation was developed as a single dose for the treatment of acute uncomplicated influenza infection, and it potentially reduces the duration of illness in participants. Although current NAIs exhibit high therapeutic efficacy against circulating influenza virus, searching for novel anti-influenza agents is continuously performed to address newly emerging or mutant strains with resistance to anti-influenza agents.

Nevertheless, a number of drug candidates have failed in the late stages of the drug development process, primarily during clinical trials. These failures are a result of either insufficient therapeutic efficacy or adverse drug reactions at therapeutic doses. Balancing between favorable bioactivity and desirable adverse effects is essential for improving the therapeutic outcome after treatment (Greene and Naven, 2009). The bioactivity of compounds is facilitated by interactions between functional groups aligning inside the molecule and target residues in the binding pocket of the drug target. Thus, insights into the structure-activity relationship are important for filling the knowledge gap during the lead optimization process. Currently, advanced computational-aided drug design approaches are employed in medicinal chemistry research, which potentially reduce costs and the amount of time spent for optimizing a set of novel compounds for pre-clinical and clinical assessments. Chemical space exploration enables the determination of important molecular substructures that contribute to bioactivity against drug targets. In combination with quantitative structure-activity relationships, the informative physicochemical properties and molecular features that are relevant to the bioactivity of compounds can be obtained for discriminating between active and inactive compounds through various machine-learning approaches.

To reduce the failure rate in the late stages of drug design and development, it is necessary to understand both important molecular substructures and informative molecular features relevant to the activity of interest. Herein, we report the application of chemical space for exploring the important structure distributions related to neuraminidase inhibitor activities and the creation of a set of simple physicochemical properties that define the preferred physicochemical properties for neuraminidase inhibition. To achieve this objective, a large data set of neuraminidase inhibitors was collected from a publicly available binding database (Liu et al., 2007). This data set provides considerable opportunity for investigating the fundamental profiles that dominate neuraminidase inhibition. Analysis of the maximum common substructures of compounds in the data set is performed to explore the chemical space of NAIs. A classification model for investigating structure-activity relationships is constructed using decision tree analysis.



## 75

## 76

77

79

80

81

(1)

88

82

83

94

84

85

22

86

87

22

88

89

..

90

91

22

92

93

**Table 1** Summary of the data set used for predicting the inhibitory activity of influenza type A and B.

Data set	Initial	Internal data set		External data set	
		Active	Inactive	Active	Inactive
Type A	410	204	124	51	31
Type B	171	54	83	13	21

## 94 Molecular descriptor generation

95 A molecular descriptor is a numerical description that represents the physicochemical properties and  
96 chemical information of compounds. The chemical structures of curated NAIs in the form of SMILES  
97 structures were converted to 3D structures using MolConverter from ChemAxon (ChemAxon Ltd.,  
98 2015b) and then subsequently converted to Gaussian input file format using Open Babel (O'Boyle et al.,  
99 2011). Geometrical optimization was performed using density functional theory (DFT) calculations  
100 at the B3LYP/6-31G(d) level as implemented in Gaussian09 (Frisch et al., 2009). In this study, low-  
101 energy conformations obtained from geometrical optimizations were used to extract thirteen easy-to-  
102 interpret molecular descriptors, consisting of six quantum chemical descriptors and seven molecular  
103 descriptors, accounting for the physicochemical properties of compounds according to our previous study  
104 (Nantasenamat et al., 2013). The obtained quantum chemical descriptors include the mean absolute charge  
105 ( $Q_m$ ), energy (E), dipole moment ( $\mu$ ), highest occupied molecular orbital (HOMO), lowest unoccupied  
106 molecular orbital (LUMO), and energy gap of the HOMO and LUMO state (HOMO-LUMO gap).  
107 Furthermore, the second sets of molecular descriptors were calculated using DRAGON 5.5 Professional  
108 (Talete srl., 2007). The obtained descriptors include the molecular weight (MW), rotatable bond number  
109 (RBN), number of rings (nCIC), number of hydrogen bond donors (nHDon), number of hydrogen bond  
110 acceptors (nHAcc), Ghose-Crippen octanol-water partition coefficient (ALogP), and topological polar  
111 surface area (TPSA).

## 112 Univariate analysis

113 As an exploratory data analysis (EDA), univariate statistical analysis was performed to investigate the  
114 different patterns and trends of individual molecular descriptors between active and inactive NAIs using 6  
115 descriptive statistical parameters: the minimum (Min), first quartile (Q1), median, mean, third quartile  
116 (Q3) and maximum (Max). In addition, statistical differences of descriptors among active and inactive  
117 NAIs were evaluated using the p-value obtained from Student's t-test (Goodman, 1999). Finally, histogram  
118 plots of the thirteen descriptors were generated using in-house R language scripts to visualize the different  
119 distributions of active and inactive NAIs.

## 120 Data splitting

121 The aforementioned non-redundant data sets were divided into internal and external sets with the Kennard-  
122 Stone sampling algorithm (Stevens, 2014) using ratios of 80% and 20%, respectively (Table 1). The  
123 internal set was subjected to full training calculations and was evaluated using a ten-fold cross-validation  
124 (10-fold CV) scheme, which was applied to confirm the reliability and robustness of the proposed models.  
125 Furthermore, the external set was used to assess the generalization ability of the model when extrapolating  
126 to unknown data samples.

## 127 Multivariate analysis

128 Decision tree (DT) is a simple, transparent and interpretable learning method that produces decision rules  
129 for the underlying data (Quinlan, 1993). Practically, the prediction task using the decision model can  
130 be easily implemented without complicated computations, and this model can also be applied in both  
131 continuous and categorical variables (Prachayasittikul et al., 2015). This algorithm has been widely used  
132 for the interpretable analysis of various tasks, such as hepatitis virus C NS5B polymerase (Nantasenamat  
133 et al., 2010), aromatase inhibitors (Nantasenamat et al., 2013; Shoombuatong et al., 2015b), dipeptidyl  
134 peptidase IV inhibitors (Shoombuatong et al., 2015a), and metabolic syndrome (Worachartcheewan et al.,  
135 2013). This study employs Weka's (Hall et al., 2009) J48 algorithm (a Java implementation of the C4.5  
136 algorithm) for constructing a predictive model for discriminating influenza virus type A and B into its  
137 class (active or inactive group). The model is constructed as a function of a set of thirteen molecular  
138 descriptors. In the J48 algorithm, the information gain is used to rank features for constructing a decision  
139 tree based on feature usage. The feature usage score can be obtained after constructing a decision tree and  
140 then counting the firing frequency of associated rules (nodes). The feature usage provides an easy way to  
141 rank and identify important features. A molecular descriptor with a high feature usage is considered to be  
142 an important feature.

143 Principal component analysis (PCA) is a tool used for analyzing data sets that possess several inter-  
144 correlated quantitative dependent variables (Prachayasittikul et al., 2015; Jolliffe, 2005). To manipulate  
145 these inter-correlated variables, PCA essentially transforms the original data into a number of principal  
146 components (PCs) or new co-ordinate axes, where the axes are located on the center of the data points.



Mathematically, PCs depends on the eigenvectors and eigenvalues of a data covariance (or correlation) matrix. The eigenvector associated with the largest eigenvalue has a direction that is identical to the first principal component (PC1), whereas the eigenvector associated with the second largest eigenvalue determines the direction of the second principal component (PC2) and so forth. In the present study, PCA was utilized in influenza virus type A and B, cooperating with the thirteen molecular descriptors to provide a better understanding of neuraminidase by using the FactoMineR (Lê et al., 2008) package of the R statistical language. Prior to PCA analysis, all data were first standardized to a comparable scale by transforming variables to zero mean and unit variance.

### Statistical assessment

Four measurements were used to evaluate the prediction performance of the proposed model, namely, accuracy (Acc), sensitivity (Sen), specificity (Spec), and Matthews correlation coefficient (MCC), which are defined by the following equations:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TP + FP} \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

where TP, TN, FP and FN are the numbers of true positive, true negative, false positive and false negative, respectively.

### Maximum common substructure analysis

The chemical substructure analysis or molecular fragment analysis was performed to analyze the properties of the NAIs expressed by molecular descriptors using LibMCS software as implemented in ChemAxon's JChem technology to identify and display the maximum common substructures of compounds in the data set (ChemAxon Ltd., 2015a). In brief, all chemical structures in SMILES format were initially converted to SDF format as an input file using MolConverter (ChemAxon Ltd., 2015b). LibMCS subsequently generated maximum common substructures present in the data set. The fragments were ranked according to structure-based hierarchical clustering algorithms, in which the bottoms of the hierarchy are the initial structure, and then the next level contains the maximum common substructures of initial molecule clusters where all molecules that share the same common structure are placed in a cluster. Active and inactive fragments were distinguished according to pIC<sub>50</sub> cut-off values of >6 and <5, respectively, and the chemical substructures were ranked according to their fragment occurrence in both the active and inactive groups of the data set.

### Functional group analysis

Functional group analysis was performed to identify the important functional groups relevant for bioactivity against neuraminidase of influenza A and B. Low-energy conformation structures of both NAI data sets were employed to calculate functional group descriptors using DRAGON 5.5 (Talet srl., 2007). A total of 154 functional group descriptors were obtained for both NAI data sets against influenza A and B. Prior to analyzing the informative descriptors, constant variables with a standard deviation (SD) of less than 0.05 were eliminated from the data set, which resulted in 59 and 46 descriptors remaining for NAIs against influenza A and B, respectively. In this study, decision tree models were constructed to distinguish active compounds from inactive compounds using the J48 algorithm implemented by Weka (Hall et al., 2009). Furthermore, the informative functional group descriptors were observed from the percentage of feature

usage as calculated by the C5.0 package of the R statistical language (Kuhn et al., 2015). Thus, rules for classifying active and inactive NAIs for both types of influenza were obtained, and the influence of functional groups relevant to bioactivity was determined from decision models. In addition, the propensity scores of functional group descriptors were calculated to reveal distribution patterns between active and inactive NAIs from both data sets. This score was calculated according to the following equation:

$$PS_{Fn(i)} = \frac{SumAc_{Fn(i)}}{SumAc} - \frac{SumInAc_{Fn(i)}}{SumInAc} \quad (6)$$

where  $PS_{Fn(i)}$  is the propensity score for the  $i^{th}$  functional group;  $SumAc_{Fn(i)}$  and  $SumInAc_{Fn(i)}$  are the total number of  $i^{th}$  functional group in the active and inactive NAI data sets, respectively;  $SumAc$  and  $SumInAc$  are the total number of all functional groups in the active and inactive NAI data sets. Finally, the propensity scores of all amino acids were normalized into the range of [0,1000] (Charoenkwan et al., 2013).

## Binding analysis

To further understand the protein-ligand interaction site, a structure-based molecular docking approach was employed in this study. Sets of 148 and 45 active NAIs against influenza type A and B, respectively, were subjected to docking with neuraminidase. In this study, the crystal structures of neuraminidase N1pdm2009 (PDB accession code 3TI4) and B (PDB accession code 1A4G) were retrieved from the Protein Data Bank (Berman et al., 2000) and were responsible for neuraminidase of influenza A and B, respectively. The proteins were initially prepared by removing water molecules and alternative side chains. Hydrogens and Gasteiger charges were added to the macromolecules, which were subsequently cleaned up by merging the charges, repairing bonds and removing non-polar hydrogens and lone pair atoms. Low-energy conformers of active NAIs obtained from the geometrical optimization process were employed to dock with the binding site of neuraminidase. Grid boxes were generated by centering on the ligand with dimensions of 40 Å 30 Å 32 Å and 40 Å 40 Å 40 Å to cover the active site of influenza type A and B neuraminidase, respectively. Molecular docking was performed using AutoDock Vina (Trott and Olson, 2010) with default parameters. The docking protocols were subsequently validated by calculating the root-mean-square deviation (RMSD) of atomic positions between co-crystallized ligand and re-binding ligand, which are laninamivir octanoate and zanamivir for PDB ID: 3TI4 and 1A4G, respectively. The protocol is accepted with an  $RMSD \leq 2.0$  Å, which was observed to be 1.153 and 1.277 Å for 3TI4 and 1A4G, respectively.

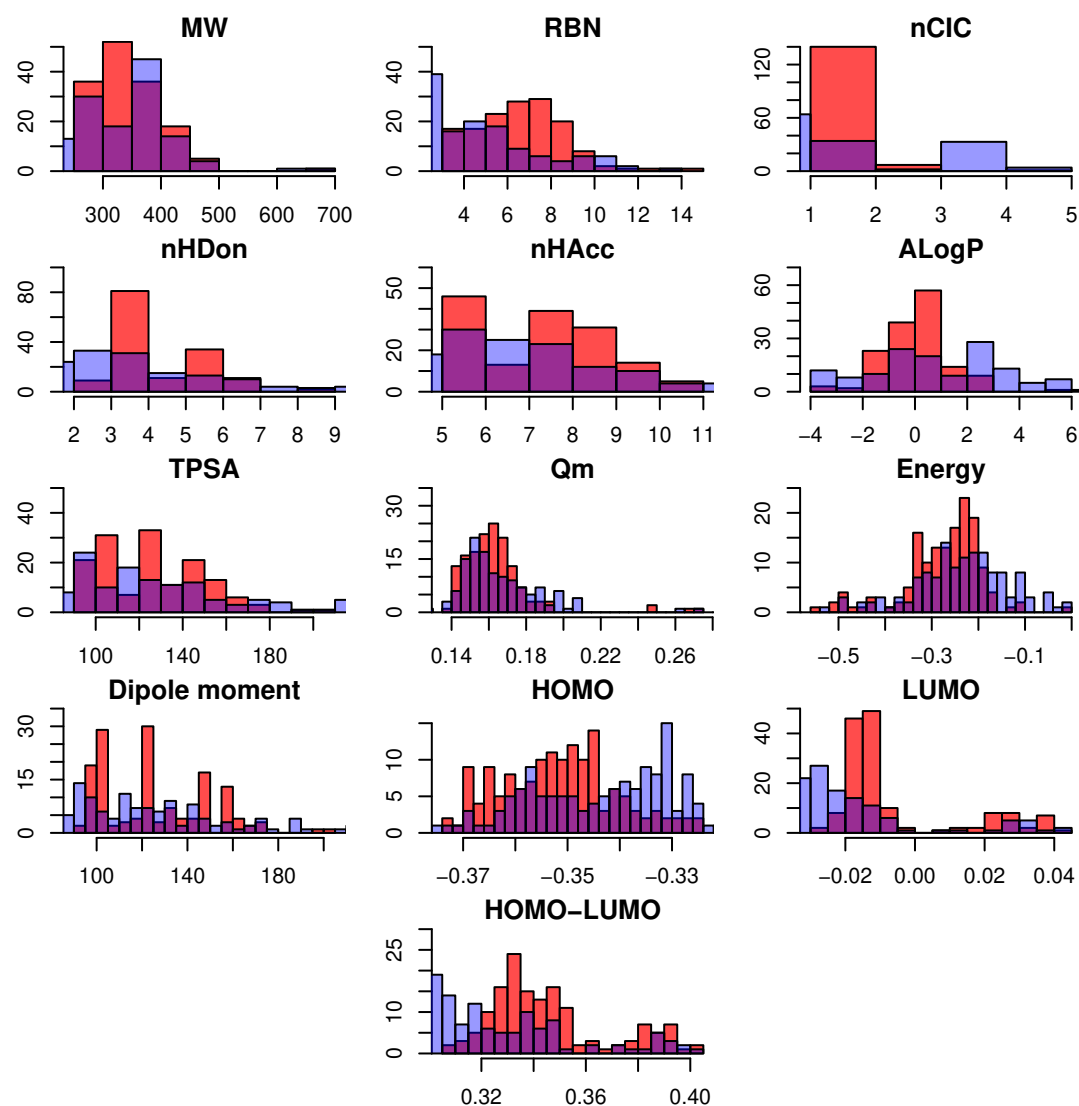
## RESULTS AND DISCUSSION

### Univariate analysis of influenza type A and B neuraminidase inhibitors

A total of 313 NAIs collected from the BindingDB consist of 285 and 131 NAIs targeting influenza type A and B neuraminidase, respectively, as shown in Table 1. Because NAIs are used to inhibit both influenza type A and type B neuraminidase, in which distinct protein structures alter the efficacy of treatment, the influenza type A and type B neuraminidase were analyzed separately to obtain a better understanding of individual pharmacokinetic properties. To determine the different characteristics between active and inactive on both influenza A and B NAIs, a univariate analysis approach based on EDA and histogram plots was used, as shown in Tables 2–5 and Figure 2. The thirteen descriptors responding to the pharmacokinetic properties of the compounds were used to provide an overview of the distribution of data values. The bioactivities of the NAIs were determined by observing the mean  $pIC_{50}$  value, which was  $5.788 \pm 2.023$  (1.30 μM) and  $5.107 \pm 1.695$  (7.80 μM) for type A and B neuraminidase, respectively. It could be observed that NAIs for influenza type A neuraminidase possessed significantly different therapeutic activity than those for type B neuraminidase with  $p < 0.05$ .

Herein, the physicochemical properties of the NAIs were used to analyze either individual descriptors or several descriptors in conjunction. The statistical analysis results showed that the NAIs exhibited good agreement with the properties of known drugs, as suggested and mentioned by Lipinski's rule of 5 for drug-like molecules (Lipinski et al., 2001). Molecular features of FDA-approved drugs were observed and used for establishing the simple rule for drug-like molecules, which generally exhibited the following molecular features: (1) MW < 500 Da, (2) LogP < 5, (3) nHDon < 5, and (4) nHAcc < 10.





**Figure 2** Histogram representing the molecular descriptors for NAIs against influenza type A. Note: Active and inactive NAIs are represented with red and blue bars, respectively, whereas the purple represents their overlap region.

Structure-activity relationships of the NAIs were further observed based on their molecular features and bioactivities. As described above, the compounds were classified as active or inactive using  $pIC_{50}$  cut-offs of  $\geq 6$  ( $IC_{50} \leq 1 \mu M$ ) and  $\leq 5$  ( $IC_{50} \geq 10 \mu M$ ), respectively; however, compounds that exhibited a  $pIC_{50}$  value in the range of 5 to 6 were not considered in this study (similar to the Data Collection section).

MW is the response to the molecular size of compounds. Following Lipinski's rule of five for drug-like molecules, MW was indicated as one of the important parameters according to Lipinski's rule of five. Statistical analysis showed that the average molecular size of active compounds for influenza type A NAIs ( $343.234 \pm 56.916$ ) was not significantly different from that of inactive compounds ( $328.415 \pm 83.823$ ) with  $p = 0.085$ , where the values of the descriptive statistics are Min = 254.4, Q1 = 300.4, Median = 332.4, Q3 = 372.3, and Max = 665.9 for active influenza type A NAIs and Min = 145.2, Q1 = 265.3, Median = 344.4, Q3 = 383.4, and Max = 665.9 for inactive influenza type A NAIs, as shown in Table 3. However, for influenza type B NAIs, the average MW of the active ( $312.466 \pm 44.641$ ) and inactive ( $357.692 \pm 76.866$ ) groups were significantly different with  $p < 0.05$ . A 6-term descriptive statistic also confirmed that the active and inactive influenza type B NAIs differed from each other, with Min = 242.3, Q1 = 284.4, Median = 300.4, Q3 = 328.5, and Max = 443.5 for the active influenza type B NAIs and Min

**Table 2** Summary of statistical analysis of active and inactive classes of influenza type A neuraminidase inhibitors.

Descriptor	Active	Inactive	<i>p</i> -value
MW	343.234 ± 56.916	328.415 ± 83.823	0.085
RBN	7.061 ± 2.183	5.248 ± 2.681	<0.001
nCIC	1.514 ± 0.655	2.109 ± 1.316	<0.001
nHDon	4.743 ± 1.257	4.321 ± 2.029	0.038
nHAcc	7.777 ± 1.502	7.204 ± 2.153	0.01
ALogP	0.049 ± 1.256	0.853 ± 2.475	<0.001
TPSA	125.205 ± 24.370	120.033 ± 37.060	0.169
$Q_m$	0.163 ± 0.020	0.179 ± 0.055	0.002
Energy	-0.274 ± 0.083	-0.246 ± 0.119	0.026
Dipole moment	4.101 ± 1.972	4.328 ± 2.003	0.336
HOMO	-0.352 ± 0.011	-0.343 ± 0.015	<0.001
LUMO	-0.006 ± 0.018	-0.021 ± 0.022	<0.001
HOMO-LUMO	0.346 ± 0.023	0.322 ± 0.030	<0.001

= 194.2, Q1 = 309.8, Median = 357.5, Q3 = 390.5, and Max = 665.9 for the inactive influenza type B NAIs, as shown in Table 5.

RBN is the number of rotatable bonds in a molecule and provides a relative measure of molecular flexibility. RBN is defined as any single bond, not in a ring, bound to a non-terminal heavy atom. Amide C–N bonds are excluded from the count because of their high rotational energy barrier. As shown in Table 2, the number of rotatable bonds in a molecule of the active group (7.061±2.183) for influenza type A NAIs is distinctly dissimilar from that of the inactive group (5.248±2.681). In the case of influenza type B NAIs, the active group (5.689±2.043) is also dissimilar from the inactive group (7.233±2.561), as shown in Table 4. Additionally, a 6-term descriptive statistic also revealed that active and inactive influenza type A and B NAIs are different. All results indicate that the number of rotatable bonds in a molecule of the active group for both influenza type A and B NAIs is significantly different from that of the inactive group at the  $p < 0.05$  level.

The number of rings (nCIC) is calculated as the cardinality of the set of independent rings known as the smallest set of smallest rings (SSSR). As shown in Tables 2 and 3, the average number of rings of the active group (1.514±0.655) of influenza type A NAIs is less than that of the inactive group (2.109±1.316). Similar to type B, the average number of rings of the active group (1.444±0.546) is not greater than that of the inactive group (1.709±0.852). A 6-term descriptive statistic confirms that the active and inactive groups of influenza type A and B NAIs are significantly different at the level of  $p < 0.05$ , where the nCIC values of influenza type A NAIs are in the range of [1.000, 5.000] and [1.000, 5.000] for active and inactive groups, respectively; in the case of influenza type B NAIs, ranges of [1.000, 3.000] and [1.000, 5.000] are obtained from active and inactive groups, respectively, as demonstrated in Tables 3 and 5.

nHDon is the descriptor responsible for the number of hydrogen bond donors in a molecule. In brief, the active group was found to possess higher mean values of nHDon than the inactive group for influenza type A NAIs, where as for influenza type B NAIs, the active group was found to possess lower mean values of nHDon than the inactive group. As shown in Tables 3 and 5, the nHDon values of influenza type A NAIs are in the ranges of [2.000, 9.000] and [1.000, 10.000] for the active and inactive groups, respectively, whereas the nHDon values of influenza type B NAIs range from [2.000, 9.000] and [2.000, 10.000] for the active and inactive groups, respectively. As shown in Figure 2, the histograms of nHDon in the active/inactive groups indicate that the distributions for influenza type A NAIs are significantly different, whereas the distributions for influenza type B NAIs are not significantly different at the  $p < 0.05$  level.

nHAcc is the descriptor responsible for the number of acceptor atoms in a molecule. Table 2 shows that the number of acceptor atoms in a molecule of the active group for influenza type A NAIs (7.777±1.502) is greater than that for inactive groups (7.204±2.153). Similar to influenza type B NAIs, the numbers of acceptor atoms in a molecule of the active (7.156±1.731) and inactive groups (8.337±1.334) are not similar to each other. The histogram plots clearly indicate that the active and inactive groups of influenza

**Table 3** Exploratory data analysis with the 6-term descriptive statistic of influenza type A neuraminidase inhibitors.

Statistics	MW	RBN	nCIC	nHDon	nHAcc	ALogP	TPSA	$Q_m$	Energy	Dipole moment	HOMO	LUMO	HOMO-LUMO
<b>Active</b>													
Min	254.4	3.00	1.00	2.00	5.00	-3.669	90.65	0.139	-0.548	0.564	-0.373	-0.03	0.308
Q1	300.4	6.00	1.00	4.00	6.00	-0.619	101.65	0.151	-0.317	2.517	-0.36	-0.016	0.331
Median	332.4	7.00	1.00	4.00	8.00	0.076	121.96	0.161	-0.256	3.791	-0.352	-0.014	0.34
Mean	343.2	7.061	1.514	4.743	7.777	0.049	125.21	0.163	-0.274	4.101	-0.352	-0.006	0.346
Q3	372.3	8.00	2.00	6.00	9.00	0.595	148.61	0.169	-0.221	5.338	-0.345	-0.009	0.351
Max	665.9	15.00	5.00	9.00	11.00	5.614	200.72	0.272	0.00	9.977	-0.326	0.041	0.405
<b>Inactive</b>													
Min	145.2	1.00	0.00	1.00	1.00	-3.807	20.23	0.115	-0.681	0.208	-0.385	-0.066	0.272
Q1	265.3	3.00	1.00	3.00	6.00	-0.858	94.14	0.152	-0.301	2.82	-0.355	-0.033	0.301
Median	344.4	5.00	2.00	4.00	7.00	0.523	114.37	0.162	-0.237	4.176	-0.341	-0.026	0.316
Mean	328.4	5.248	2.109	4.321	7.204	0.853	120.03	0.179	-0.246	4.328	-0.343	-0.021	0.322
Q3	383.4	6.00	4.00	5.00	8.00	2.599	140.57	0.182	-0.172	5.459	-0.332	-0.015	0.339
Max	665.9	14.00	5.00	10.00	14.00	6.834	218.97	0.434	0.00	10.561	-0.311	0.044	0.403

**Table 4** Summary of statistical analysis of active and inactive classes of influenza type B neuraminidase inhibitors.

Descriptor	Active	Inactive	<i>p</i> -value
MW	312.466 ± 44.641	357.692 ± 76.866	<0.001
RBN	5.689 ± 2.043	7.233 ± 2.561	<0.001
nCIC	1.444 ± 0.546	1.709 ± 0.852	0.033
nHDon	4.578 ± 1.340	4.849 ± 1.561	0.302
nHAcc	7.156 ± 1.731	8.337 ± 1.334	<0.001
ALogP	-0.128 ± 1.291	-0.017 ± 1.602	0.67
TPSA	114.592 ± 27.661	134.168 ± 25.007	<0.001
$Q_m$	0.158 ± 0.013	0.168 ± 0.018	<0.001
Energy	-0.276 ± 0.099	-0.249 ± 0.089	0.129
Dipole moment	3.831 ± 1.601	4.364 ± 2.011	0.101
HOMO	-0.351 ± 0.009	-0.350 ± 0.013	0.573
LUMO	-0.013 ± 0.010	-0.003 ± 0.023	0.002
HOMO-LUMO	0.339 ± 0.014	0.347 ± 0.029	0.031

type A and B NAIs are quite dissimilar, as shown in Figures 2 and 3. In summary, all of these results indicated that the number of acceptor atoms in a molecule between active and inactive groups of influenza type A and B NAIs were significantly different at the  $p < 0.05$  level.

ALogP is a computational method for estimating the 1-octanol/water partition coefficient (logP), which is a well-known measure of molecular hydrophobicity also known as lipophilicity. As shown in Figures 2 and 3, the histogram of ALogP of influenza type B has a greater overlapping region (purple) than that of type A. Tables 2 and 4 demonstrate that these results were compatible with the average value, with  $0.049 \pm 1.256$  and  $0.853 \pm 2.475$  for active and inactive influenza type A NAIs, respectively, whereas for influenza type B NAIs, the average values of active and inactive are  $-0.128 \pm 1.291$  and  $-0.017 \pm 1.602$ , respectively. In summary, for ALogP, the active group of influenza type A NAIs is significantly different from the inactive group at the  $p < 0.05$  level, as shown in Table 2, whereas in the case of influenza type B NAIs, the active is quite similar.

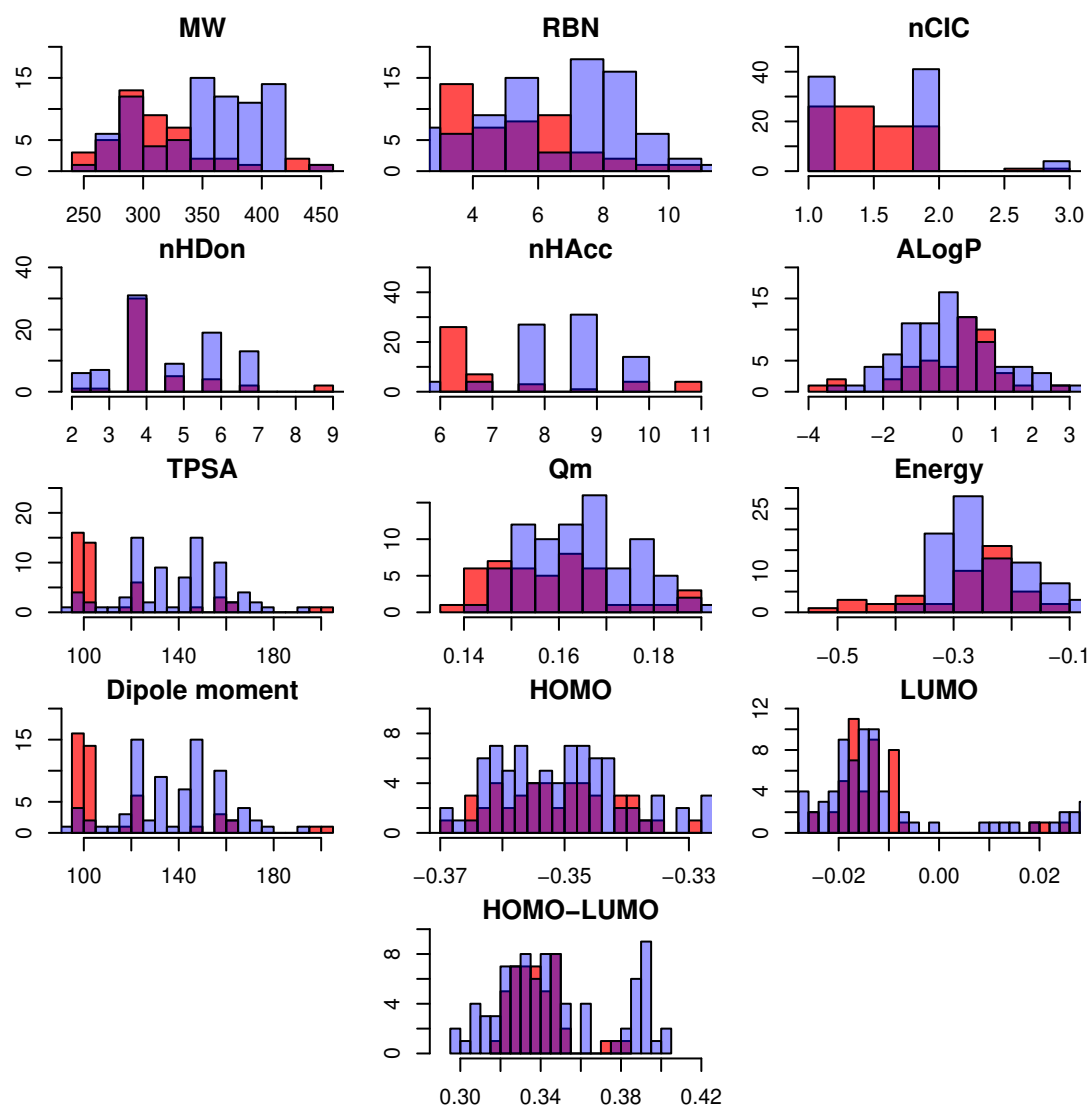
TPSA describes the contribution of polar atoms to the molecular charge based on an empirical measurement of the polar surface area of a molecule. Tables 2 and 4 show that the statistical analysis of influenza A NAI is not significant at the  $p < 0.05$  level, whereas statistically significant results can be observed among the active and inactive groups of influenza type B NAIs. The corresponding TPSA values were  $125.205 \pm 24.370$  (active) and  $120.033 \pm 37.060$  (inactive) for influenza A NAIs, whereas the TPSA values of influenza B NAIs were  $114/592 \pm 27.661$  (active) and  $134.168 \pm 25.007$  (inactive). A 6-term descriptive statistic is compatible with the statistically different results of both influenza A and B NAIs, where Min = 90.650/20.230, Q1 = 101.650/94.140, Median = 121.960/114.370, Q3 = 148.610/140.570, and Max = 200.720/218.970 for actives/inactives of influenza A NAIs, and in the case of influenza B NAIs, Min = 95.660/69.640, Q1 = 95.660/121.960, Median = 101.650/135.600, Q3 = 121.960/148.610, and Max = 200.720/192.700.

The mean absolute charge ( $Q_m$ ) is the response to the global measurement of molecular charge. The histogram plot showed different distributions of mean absolute charge among influenza type A and type B NAIs. Moreover, the  $Q_m$  here exhibited a distinct mean absolute charge of compounds with  $0.171 \pm 0.042$  and  $0.165 \pm 0.017$  for influenza type A and type B NAIs, respectively. This study suggested that the inactive group had higher  $Q_m$  values compared to the active group, as shown in Tables 2 and 4. The  $Q_m$  values for the active and inactive influenza type A NAIs were  $0.163 \pm 0.020$  and  $0.179 \pm 0.055$ , respectively, whereas those for the influenza type B NAIs were  $0.158 \pm 0.013$  and  $0.168 \pm 0.018$  for the active and inactive groups, respectively. Statistical analysis indicated that the active and inactive compounds for influenza type A and B exhibited significant differences in their charges at the  $p < 0.05$  level.

Energy is the response to the summation of the atomic energy. Overall, insignificant differences in energy among inhibitors of influenza type A ( $-0.260 \pm 0.103$ ) and type B ( $-0.258 \pm 0.093$ ) neuraminidase were observed at the  $p = 0.823$  level. It was found that the active group ( $-0.274 \pm 0.083$ ) had a slightly

**Table 5** Exploratory data analysis with the 6-term descriptive statistic of influenza type B neuraminidase inhibitors.

Statistics	MW	RBN	nCIC	nHDon	nHAcc	ALogP	TPSA	$Q_m$	Energy	Dipole moment	HOMO	LUMO	HOMO-LUMO
<b>Active</b>													
Min	242.3	3.00	1.00	2.00	6.00	-3.669	95.66	0.139	-0.548	1.407	-0.368	-0.025	0.315
Q1	284.4	4.00	1.00	4.00	6.00	-0.722	95.66	0.148	-0.302	2.78	-0.358	-0.017	0.329
Median	300.4	6.00	1.00	4.00	6.00	0.18	101.65	0.157	-0.248	3.468	-0.353	-0.015	0.337
Mean	312.5	5.689	1.444	4.578	7.156	-0.128	114.59	0.158	-0.276	3.831	-0.351	-0.013	0.339
Q3	328.5	7.00	2.00	5.00	8.00	0.585	121.96	0.165	-0.207	4.811	-0.346	-0.01	0.345
Max	443.5	11.00	3.00	9.00	11.00	2.727	200.72	0.189	-0.135	8.19	-0.329	0.024	0.383
<b>Inactive</b>													
Min	194.2	2.00	1.00	2.00	5.00	-3.389	69.64	0.145	-0.679	0.564	-0.384	-0.033	0.297
Q1	309.8	5.25	1.00	4.00	8.00	-1.052	121.96	0.156	-0.301	2.94	-0.359	-0.019	0.326
Median	357.5	8.00	2.00	4.00	9.00	-0.205	135.6	0.166	-0.264	4.006	-0.35	-0.014	0.342
Mean	357.7	7.233	1.709	4.849	8.337	-0.017	134.17	0.168	-0.249	4.364	-0.35	-0.003	0.347
Q3	390.5	9.00	2.00	6.00	9.00	0.615	148.61	0.176	-0.195	5.782	-0.344	0.017	0.363
Max	665.9	14.00	5.00	10.00	10.00	5.614	192.7	0.272	0.00	9.791	-0.314	0.041	0.405



**Figure 3** Histogram representing the molecular descriptors for NAIs against influenza type B. Note: Active and inactive NAIs are represented with red and blue bars, respectively, whereas the purple represents their overlap region.

higher energy than the inactive group ( $-0.246 \pm 0.119$ ) for influenza type A. However, the values of ( $-0.276 \pm 0.099$ ) and ( $-0.249 \pm 0.089$ ), which were observed in the active and inactive groups, respectively, of influenza B NAIs exhibited insignificant differences at the  $p < 0.05$  ( $p = 0.129$ ) level. Tables 3 and 5 also summarize the exploratory data analyses, which are consistent with the statistical analyses with Min =  $-0.548/-0.681$ , Q1 =  $-0.317/-0.301$ , Median =  $-0.256/-0.237$ , Q3 =  $-0.221/-0.172$ , and Max =  $-0.135/0.000$  for actives/inactives of influenza A NAIs, and in the case of influenza B NAIs, Min =  $-0.548/-0.679$ , Q1 =  $-0.302/-0.301$ , Median =  $-0.248/-0.264$ , Q3 =  $-0.207/-0.195$ , and Max =  $-0.135/0.000$ .

Dipole moment is the response to the asymmetric distribution of charge in the molecule. A high dipole moment value indicates a high charge distribution and vice versa. The overviews of the dipole moments of influenza type A ( $4.210 \pm 1.987$ ) and B ( $4.181 \pm 1.891$ ) NAIs were not significantly different at the  $p < 0.05$  ( $p = 0.737$ ) level. Statistical analysis was also performed to reveal the informative pattern of type A and B between the active and inactive groups. Remarkably, the dipole moments of both influenza type A and B NAIs are not significantly different, with  $p = 0.336$  and  $p = 0.537$ , respectively. Remarkably, the exploratory data analyses clearly indicate that the dipole moments of both influenza type A and B NAIs between active and inactive groups are quite similar, as shown in Tables 3 and 5.



The highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) are the highest- and lowest-energy molecular orbitals that are occupied and unoccupied by electrons, respectively. For the mean values of the HOMO of the active and inactive groups of influenza type A and B NAIs, it could be observed that the HOMO value of the active group is significantly different from that of the inactive group at the  $p < 0.05$  level. Moreover, in the case of influenza type B NAIs, the HOMO value of the active group is not a statistically significant result ( $p = 0.573$ ) at the  $p < 0.05$  level, as summarized in Tables 2 and 4. However, the mean values of the LUMO of influenza type A and B NAIs present statistically significant results between the active and inactive groups at the  $p < 0.05$  level, as summarized in Tables 2 and 4. Additionally, a 6-term descriptive statistic also confirmed that the active and inactive influenza type A and B NAIs are different.

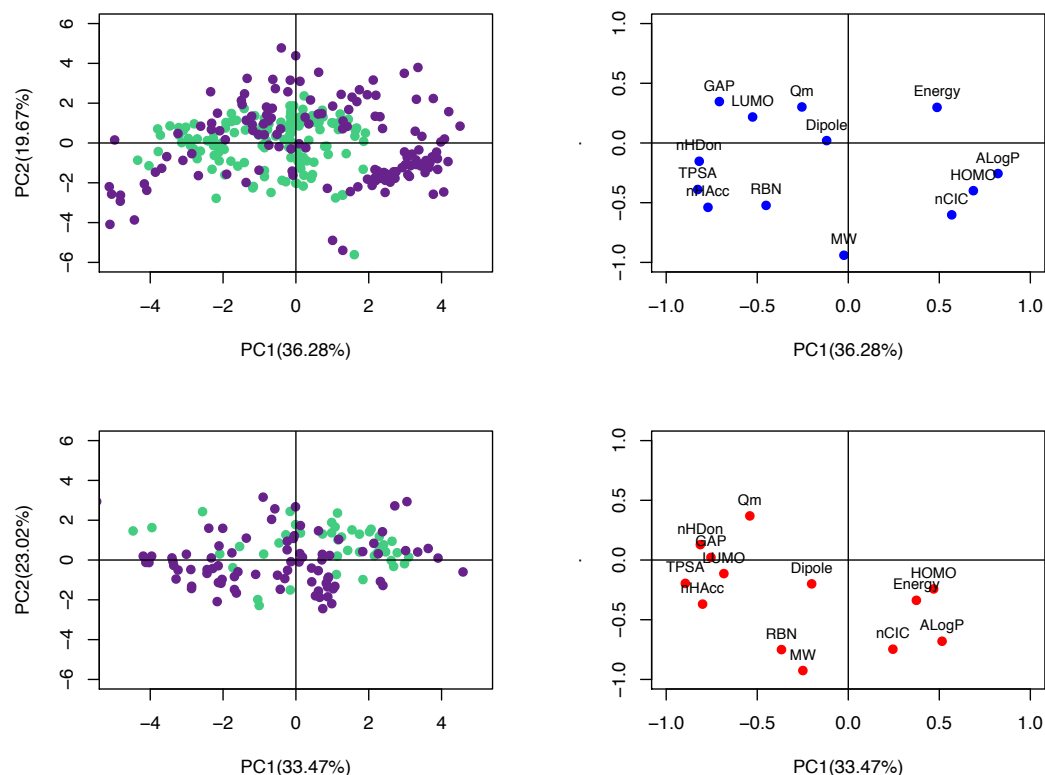
In quantum chemistry, the energy gap between the HOMO and LUMO (HOMO-LUMO) has been used to measure the kinetic stability and chemical reactivity of molecules. The energy of the HOMO is associated with ionization potential (ability to donate electrons), whereas the LUMO is responsible for electron affinity (ability to accept electrons). A small energy gap between these two states is related to a low kinetic stability and provides high chemical reactivity and vice versa (32). The histogram plot shows a slightly different pattern of distribution of the NAIs for influenza type A and B between the active and inactive groups (Figures 2 and 3, respectively). The overviews of the HOMO-LUMO values of influenza type A ( $4.210 \pm 1.987$ ) and B ( $4.181 \pm 1.891$ ) NAIs were significantly different at the  $p = 0.737$  level. For analysis of influenza type A and B NAIs, the HOMO-LUMO value of the active group is shown with the statistically significant results at the  $p < 0.05$  level for both type A and B, as summarized in Tables 2 and 4.

In summary, all of these results indicated that nearly all of the 13 descriptors were significantly different between the active and inactive groups of influenza type A NAIs at the level of  $p < 0.05$ , except for MW ( $p = 0.085$ ), TPSA ( $p = 0.169$ ) and dipole moment ( $p = 0.0336$ ). With the exception of the MW, TPSA and dipole moment descriptors, the remaining descriptors are significantly different for the active and inactive groups of influenza type A NAIs and are efficient for discrimination. Similar to influenza type B NAIs, with the exception of the nHDon ( $p = 0.302$ ), ALogP ( $p = 0.670$ ), energy ( $p = 0.129$ ), dipole moment ( $p = 0.573$ ), and HOMO ( $p = 0.085$ ) descriptors, the remaining descriptors are efficient for discrimination. However, in practice, the compounds used for treating influenza type B are the same compounds used to develop treatments for influenza type A. Thus, the univariate analysis of compound properties cannot provide the important relationships among molecular features that affect the treatment efficacy. Herein, multivariate analysis and classification of structure-activity relationships were performed to investigate such important features using principal component analysis (PCA) and decision tree analysis.

### PCA analysis of influenza type A and B neuraminidase inhibitors

The PCA method was used to transform the data set of influenza type A (Figure 4 (Upper)) and B (Figure 4 (Lower)) to a few principal components (PCs), in which significant variances among variables are revealed by the eigenvalues or variance. PC1 had the highest variance in the data of influenza type A NAIs, with 36.28% of the original variance for the active and inactive groups. Meanwhile, PC2 and PC3 provided the highest second and third variances, with 19.67% and 13.29% of the original variance for the active and inactive groups, respectively. In summary, the first three PCs indicated that the amount of cumulative variation of these PCs is as high as almost 70% of the original variance, which was sufficiently informative data for further analysis. Figure 4 (A) shows that the descriptors of nCIC, HOMO and ALogP are the main descriptors responsible for the inactive group. MW had a considerable influence on the segregation of two samples in the inactive group and one sample in the active group from the remaining data points, whereas nHDon, TPSA, nHAcc, and RBN primarily contributed to the segregation of eight samples of the inactive group from the other data. The results still showed that the descriptors of  $Q_m$ , energy, ALogP, HOMO, and nCIC may be the main descriptors allowing for the discrimination of active and inactive groups.

In the case of the data for influenza type B NAIs, PC1 had the highest variance in the data of influenza type B NAIs, with 33.47% of the original variance for the active and inactive groups. Meanwhile, PC2 and PC3 provided the highest second and third variances, with 23.02% and 23.04% of the original variance for the active and inactive groups, respectively. Similar to type A, the first three PCs still indicated that the amount of cumulative variation of these PCs are as high as almost 70% of the original variance. Figure 4



**Figure 4** PCA score plots and loading plots of NAIs against influenza type A (A) and B (B). Left and right panels are the score and loading plots, respectively, and the active and inactive compounds are represented by green and purple dots in the score plots.

(B) shows that almost thirteen descriptors were considered sufficient for describing the active and inactive groups, except for RBN and MW. Note that using these descriptors may improve the performance of a predictive model for discriminating influenza type B into the active or inactive group.

### Prediction of inhibitory activity against neuraminidase from influenza A and B

An interpretable predictive model is more useful for providing insights into the basis of the biological and chemical properties of influenza A and B NAIs. Therefore, in this study, a QSAR model based on the J48 algorithm is presented for discriminating between active/inactive groups of influenza A and B NAIs. Each compound was calculated as an M-dimensional vector, where M=13. To construct a predictive model, the J48 algorithm was applied using the encoded compounds from the internal sets. Moreover, to evaluate the ability of our proposed QSAR model, two different experiments were performed: one experiment was performed on the full training data, and the other experiment was evaluated using a 10-fold CV procedure, as shown in Table 6. The CV procedure was performed by first partitioning the data into 10 equally sized segments or folds; then, 9 folds were used as the training data, while the remaining fold was used for validation. Finally, the results were averaged across the 10 experiments. Four measurements were used to assess the performance of the QSAR models, namely, accuracy (Acc), sensitivity (Sen), specificity (Spec), and the Matthews correlation coefficient (MCC).

Table 6 demonstrates that using the sets of thirteen descriptors provides promising results with an accuracy of 87.50%, sensitivity of 93.63%, specificity of 77.42%, and MCC value of 0.731 for influenza type A NAIs, whereas these descriptor sets also perform well for influenza type B NAIs with an accuracy of 89.78%, sensitivity of 87.04%, specificity of 91.57%, and MCC value of 0.731. As shown in Table 1, the used data set is not balanced because the number of positive samples (active group) is larger than that of negative samples (inactive group). Therefore, the sensitivity accuracy is considerably greater than the specificity accuracy (for influenza type A NAIs). To address this problem, the original data set

should first be balanced between the active and inactive groups. In addition, to assess the reliability of the predictive model on unknown data, an external set was considered. Table 6 shows that our proposed model still performs well for predicting influenza type A NAIs with an accuracy of 82.93%, sensitivity of 86.27%, specificity of 77.72% and MCC value of 0.637, while the performance for predicting influenza type B NAIs is acceptable with an accuracy of 70.59%, sensitivity of 76.92%, specificity of 66.67% and MCC value of 0.424. It was well recognized that a decision tree-based classifier utilized the estimated threshold to predict a sample. Thus, it was not surprising that the prediction result of type B for external validation was lower than 10-CV. However, our proposed model aims to maximize both the simplicity and interpretability of the classification method.

Molecular descriptors play an important role in representing the physicochemical properties of compounds. Identifying informative molecular descriptors will provide insights into the underlying mechanism of influenza type A and B NAIs. The feature importance for molecular descriptors is shown in Figure 5. The feature with the largest value of descriptor usage is the most important. Figure 5(a) shows that the top-three informative descriptors of influenza type A NAIs are LUMO, nHDon and nHAcc. Moreover, Figure 5(b) shows that the top-three informative descriptors of influenza type B NAI are nHAcc, nHAcc and HOMO-LUMO.

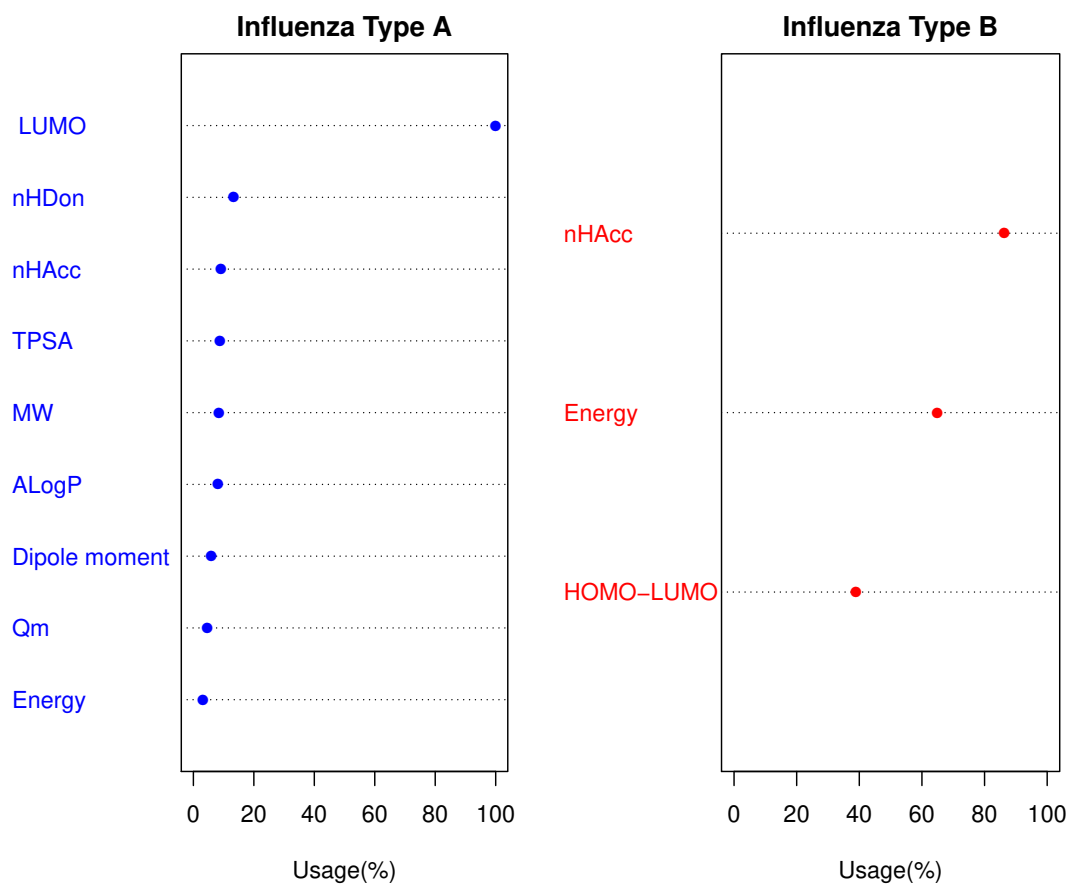
### Maximum common substructure analysis

The molecular substructure analysis revealed the important molecular fragments that facilitate the biological activity against influenza neuraminidase. The top-ranking fragments for both active and inactive NAIs are indicated in Tables 7 and 8 for influenza type A and in Tables 9 and 10 for influenza type B, respectively. The top-five fragments sorted by fragment occurrence, which resulted in high or low activity, correspond to common structures found in the chemical space of the NAIs. The results of the top-five active fragments indicated that cyclohexene-based, dihydropyran-based and cyclopentane-based fragments are relevant to inhibitory activity against influenza neuraminidase, in which these six- and five-membered non-aromatic rings possess a marginal ligand-binding conformation comparable to the tetrahydropyran ring of the sialic acid substrate of influenza neuraminidase.

The top-ranked common substructure was a cyclohexene-based moiety, which can be found in the current drug of choice for influenza treatment: oseltamivir. This drug was developed to lower the polarity effect of the dihydropyran scaffold of the first-generation NAIs, which led to the low bioavailability observed in zanamivir. Initially, zanamivir was developed based on a dihydropyran scaffold and exerts good inhibitory activity against influenza neuraminidase (Meindl et al., 1974; von Itzstein et al., 1993), which became the first approved NAI for use as a therapeutic agent against the influenza virus. Structure-based drug design based on the availability of N2 sialidase X-ray co-crystal structure with  $\alpha$ -Neu5Ac and Neu5Ac2en (Varghese et al., 1992) was used as the guideline for the development of novel NAIs. In silico analysis of enzyme active sites revealed energetically favorable interactions of amino acid residues in the active site and various functional group probes, such as carboxylates, amines, methyl groups and phosphates (von Itzstein et al., 1996). The molecular structure overlay of predicted favorable functional groups against co-crystal structure of N2 sialidase and Neu5Ac2en, as template molecules, suggest that substituting the C-4 hydroxyl group of the template with amino and guanidino groups should improve the binding affinity with the N2 active site. As a result of amino substitution at the C4 hydroxyl group, the binding affinity is enhanced by the formation of a salt bridge between the amino group and E199

**Table 6** Summary of prediction results from decision tree analysis of influenza A and B neuraminidase inhibitors.

Influenza	Performed with	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Type A	full training data	96.95	97.55	95.97	0.935
	10-CV	87.50	93.63	77.42	0.731
	Testing set	82.93	86.27	77.42	0.637
Type B	full training data	97.81	98.15	97.59	0.954
	10-CV	89.78	87.04	91.57	0.786
	Testing set	70.59	76.92	66.67	0.424

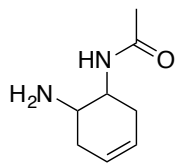
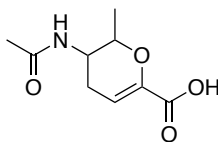
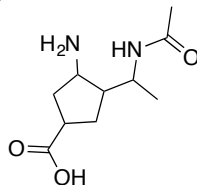
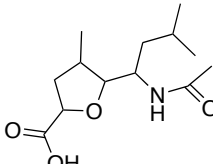
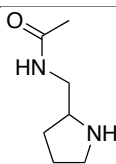


**Figure 5** Plots of the descriptor usage derived from decision tree. The descriptor with the largest percentage of descriptor usage is the most important.

residue, whereas guanidino substitution interacts with E119 and E227 via its terminal nitrogen (von Itzstein et al., 1993, 1996). Nevertheless, this acid-based inhibitor processed high polarity due to the ring oxygen and polar glycerol side chain, resulting in low bioavailability. Thus, this drug was considered to be administered by inhalation, which is difficult to provide in some patients, particularly children. The development of orally administrated NAIs was required to overcome this problem.

As previously mentioned, the polarity of dihydropyran-based NAIs affects their pharmacokinetic properties and the route of administration. To reduce the polarity effect of the dihydropyran scaffold, scaffold hopping was employed to identify appropriate molecular scaffolds that would exert desirable properties. A cyclohexene scaffold was used to replace the ring oxygen, which was previously reported to be a non-essential moiety required for neuraminidase inhibition (Taylor and von Itzstein, 1994). Replacing dihydropyran with a cyclohexene ring in which the double bond position is similar to the sialosyl transition state provided significantly higher inhibitory activity (Kim et al., 1997). Moreover, the glycerol side chain is also considered to be a main source of polarity due to its high number of oxygen atoms. The modification of the hydrophilic glycerol side chain with a 3-pentyl ether side chain based on the structure-activity relationship study led to the development of GS 4071, which was subsequently named oseltamivir carboxylate, a potent sialidase inhibitor. As a result of introducing the 3-pentyl ether side chain, the binding interaction is reorganized by reorientation of E276 from this side chain to form a salt bridge with R224, leading to the generation of a substantial hydrophobic patch, which increases the binding affinity with the ligand's hydrophobic side chain (Itzstein and Thomson, 2009). Elimination of the oxygen atom in combination with functional group modification led to lower polarity and increased the bioavailability of molecules. Thus, the second NAI was developed and consequently approved, named oseltamivir, which is currently used as a drug of choice for treating influenza. In addition, the successful development of cyclohexene-based NAIs results in the generation of extensive studies for developing novel NAIs using

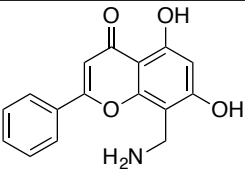
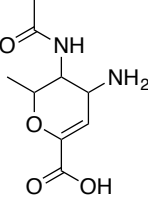
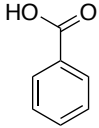
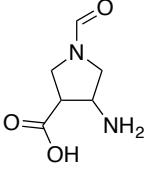
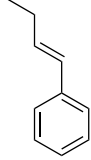
**Table 7** Summary of the top-five maximum common substructures in the active set of NAIs against influenza A.

Rank	IUPAC Name	Substructure	Substructure Occurrence
1	N-(6-aminocyclohex-3-en-1-yl)acetamide		63
2	3-acetamido-2-methyl-3,4-dihydro-2H-pyran-6-carboxylic acid		53
3	3-amino-4-(1-acetamidoethyl) cyclopentane-1-carboxylic acid		27
4	5-(1-acetamido-3-methylbutyl)-4-methyl-oxolane-2-carboxylic acid		2
5	N-[(pyrrolidin-2-yl)methyl]acetamide		2

this scaffold.

Recently, the cyclopentane scaffold in furanose was found to possess an inhibitory effect against influenza NA as strongly as the lead compound of sialidase inhibitor, called DANA. The report on inhibitory activity by furanose revealed the potential of cyclopentane as a novel scaffold for the development of NAIs (Yamamoto et al., 1992). Structure-based analysis of the cyclopentane scaffold using protein crystal structure information indicates a distinct binding mode, in which the cyclopentane ring re-organized the functional groups of NAI to interact with amino acid residues inside the binding pocket of influenza neuraminidase (Stoll et al., 2003). This evidence revealed an opportunity for introducing NAIs with novel scaffolds. The most recently approved NAI, named peramivir, was developed based on a five-membered ring scaffold. A set of novel NAIs with five-membered ring scaffolds were synthesized using cyclopentane derivatives incorporating three functional group substitutions of zanamivir, which included carboxylate group, C5-acetamido group, and C4-guanidino group, arranged in all expected positions inside the N9 active site. The functional group binding with the negatively charged area in the active site, which previously interacted with the C4 hydroxyl group of Neu5Ac2en, was designed to replace with a guanidino group as similarly observed in zanamivir. The addition of n-butyl was designed to interact with the hydrophobic region, which was previously occupied by the glycerol side chain of Neu5Ac2en. The binding interaction was confirmed by co-crystallization with N9 sialidase, and the crystal structure indicates that the binding interactions are comparable with those of zanamivir (Babu

**Table 8** Summary of the top-five maximum common substructures in the inactive set of NAIs against influenza A.

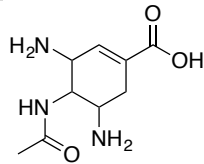
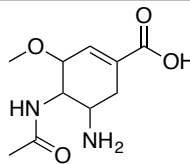
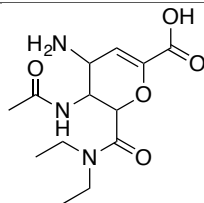
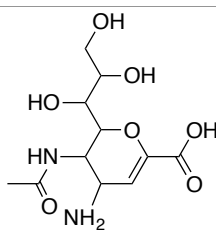
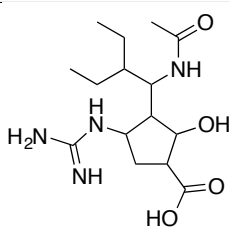
Rank	IUPAC Name	Substructure	Substructure Occurrence
1	8-(aminomethyl)-5,7-dihydroxy-2-phenyl-4H-chromen-4-one		35
2	4-amino-3-acetamido-2-methyl-3,4-dihydro-2H-pyran-6-carboxylic acid		32
3	Benzoic acid		29
4	4-amino-1-formylpyrrolidine-3-carboxylic acid		11
5	[(1E)-but-1-en-1-yl]benzene		7

et al., 2000).

Nevertheless, some of the molecular substructures that were present in the active group of NAIs, such as 3-acetamido-2-methyl-3,4-dihydro-2H-pyran-6-carboxylic acid and 5-amino-4-acetamidocyclohex-1-ene-1-carboxylic acid, can be found in the inactive group of influenza type A and B neuraminidase, respectively. Note that the inhibitory activities against influenza neuraminidase are facilitated by additional factors from both protein and ligand sides. From the protein perspective, the neuraminidase share approximately 90% structural homology in the same subtype, whereas the homology between subtypes is lower, 50% and 30%, between influenza type A and B (Shtyrya et al., 2009). The distinct structural homology affects the conformation of catalytic residues inside the catalytic pocket, resulting in different fitness binding of ligands. On the other hand, the composition of the ligand and properties affect the efficiency of the binding interaction. These factors are frequently observed by the type and position of functional groups lying in molecules, which are the crucial part for interacting with the target enzyme for inhibition. The overall size of the molecules and the molecular conformations are also important for binding with the enzyme because the binding pocket has a unique geometrical conformer that limits the shape and electrostatic properties of the target molecules. In addition, the drug-like properties of the ligand also facilitate pharmacokinetics and pharmacodynamics of ligands to reach their target and



**Table 9** Summary of the top-five maximum common substructures in the active set of NAIs against influenza B.

Rank	IUPAC Name	Substructure	Substructure Occurrence
1	3,5-diamino-4-acetamidocyclohex-1-ene-1-carboxylic acid		21
2	5-amino-4-acetamido-3-methoxycyclohex-1-ene-1-carboxylic acid		10
3	4-amino-2-(diethylcarbamoyl)-3-acetamido-3,4-dihydro-2H-pyran-6-carboxylic acid		8
4	4-amino-3-acetamido-2-(1,2,3-trihydroxypropyl)-3,4-dihydro-2H-pyran-6-carboxylic acid		3
5	4-carbamimidamido-3-(1-acetamido-2-ethylbutyl)-2-hydroxycyclopentane-1-carboxylic acid		1

generate desirable bioactivity for therapeutic purposes.

### Functional group analysis

Analyses of the functional group compositions among NAIs against both influenza type A and B were performed to observe the propensity pattern of functional group descriptors from active and inactive NAIs. The compositions of functional groups inside NAI molecules were generated from low-energy conformational structures of 285 and 131 NAIs against influenza A and B, respectively. A set of 154 functional group descriptors were obtained and consequently preprocessed by removing constant variables from each data set. As a result, 59 and 46 descriptors remained for data sets of NAIs against influenza A and B, respectively. The remaining descriptors were used to calculate the propensity score, which indicated the characteristics of NAIs regarding to their functional group compositions, as summarized in Tables and .

For NAIs against the influenza A data set, the number of total secondary  $sp^3$  carbon (nCs), number of ring secondary  $sp^3$  carbon (nCrS), number of terminal primary  $sp^3$  carbon (nCp), number of secondary

**Table 10** Summary of the top-five maximum common substructures in the inactive set of NAIs against influenza B.

Rank	IUPAC Name	Substructure	Substructure Occurrence
1	3-amino-2-methyl-3,4-dihydro-2H-pyran-6-carboxylic acid		42
2	3-carbamimidamido-4-[carbamoyl-(acetamido)methyl]cyclopentane-1-carboxylic acid		24
3	4-aminobenzoic acid		14
4	5-amino-6-acetamido-1-(2-ethylbutanoyl)-1,4,5,6-tetrahydropyridazine-3-carboxylic acid		2
5	5-amino-4-acetamidocyclohex-1-ene-1-carboxylic acid		2

aliphatic amides (nRCONHR) and number of aliphatic carboxylic acids (nRCOOH) were the top-ranked functional group descriptors abundant in active NAIs, whereas the number of aromatic  $sp^2$  carbon (nCar), number of substituted benzene  $sp^2$  carbon (nC<sub>b</sub>-), number of unsubstituted benzene  $sp^2$  carbon (nC<sub>b</sub>H), number of aromatic hydroxyls (nArOH) and number of aromatic ketones (nArCO) were the top-ranked functional group descriptors abundant in inactive NAIs. In addition, the correlation coefficient (R) between the difference of functional group compositions among active and inactive NAIs and the propensity score was 1.00. The high correlation coefficient (R) indicates that this propensity score of functional group descriptors can be distinguished between active and inactive NAIs for influenza A.

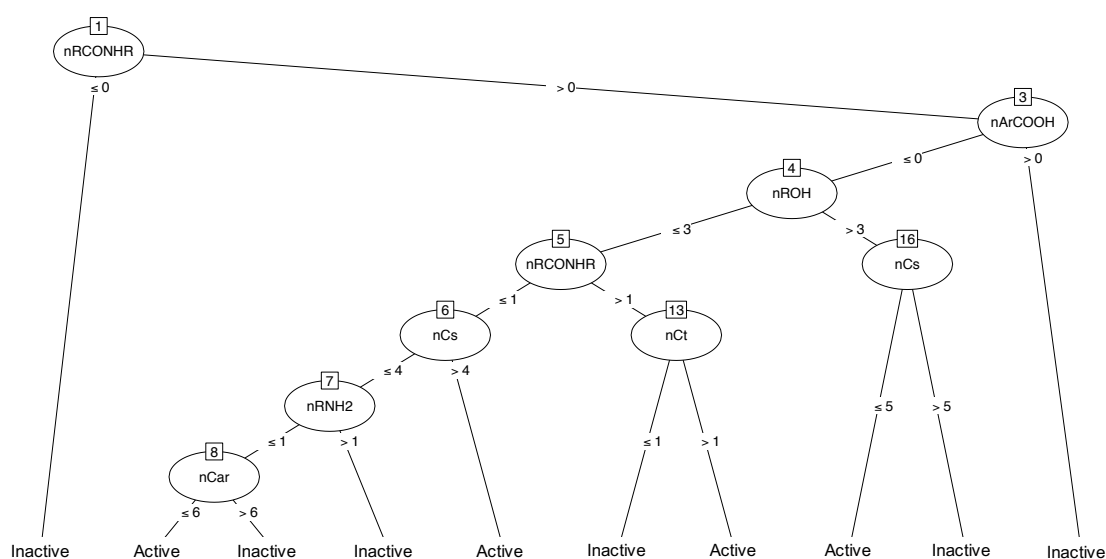
For NAIs against the influenza B data set, the number of ring secondary  $sp^3$  carbon (nCr<sub>s</sub>), number of total secondary  $sp^3$  carbon (nCs), number of non-aromatic conjugated  $sp^2$  carbon (nC<sub>conj</sub>), number of aliphatic tertiary  $sp^2$  carbon (nR=Ct) and number of aliphatic tertiary amines (nRNR<sub>2</sub>) were the majority found in the active NAIs, whereas the number of aromatic  $sp^2$  carbon (nCar), number of unsubstituted benzene  $sp^2$  carbon (nC<sub>b</sub>H), number of substituted benzene  $sp^2$  carbon (nC<sub>b</sub>-), number of total tertiary  $sp^3$  carbon (nC<sub>t</sub>) and number of ring tertiary  $sp^3$  carbon (nC<sub>rt</sub>) were mainly observed in the inactive NAIs. Furthermore, the correlation coefficient (R) of propensity score against different functional group

compositions between active and inactive classes of NAIs against influenza type B was 1.00, which indicated that the propensity score can be used to discriminate between active and inactive classes of influenza B NAIs.

To obtain a deeper understanding, decision tree models were used to construct classification models for categorizing active and inactive NAIs against both types of influenza virus. The classification models were separately generated using Quinlan's C5.0 algorithm as implemented in the C5.0 package of the R statistical language (Kuhn et al., 2015), and the predictive performances were additionally evaluated by applying 10-fold CV to the models. The CV was performed by initially splitting the data into ten equal-sized partitions, called folds. Consequently, nine folds were deployed as a training set, whereas the remaining fold was used as a validation set. The trials were continuously performed until each fold was chosen as a validation set. Finally, the predictive performances of the 10-CV were averaged through ten experiments, in which the predictive performance of the decision tree model was indicated by the percentage of accuracy, sensitivity and specificity and by the Matthews correlation coefficient (MCC).

The classification model for NAIs against influenza A exhibited excellent predictive performance, with 91.58% accuracy, 88.27% sensitivity, 95.93% specificity and 0.835 MCC. Functional group descriptors involved in this model include the number of aliphatic secondary amides (nRCONHR), number of aromatic carboxylic acids (nArCOOH), number of hydroxyl groups (nROH), number of total secondary sp<sup>3</sup> carbon (nCt), number of aliphatic primary amines (nRNH<sub>2</sub>), number of aromatic sp<sup>2</sup> carbon (nCar) and number of total tertiary sp<sup>3</sup> carbon (nCt), which are ranked by percent of descriptor usage of 100.0%, 70.2%, 65.3%, 60.0%, 12.6%, 11.9%, 5.3%, respectively. The most important functional group descriptor indicated by the largest value of descriptor usage was the number of aliphatic secondary amides (nRCONHR), which was also a root node of the model. Note that all active NAIs possessed a nRCONHR ≥ 1 in their molecules, whereas compounds lacking nRCONHR were classified as inactive NAIs according to the decision tree model summarized in Figure 6. Supported by the propensity score, nRCONHR is located in the top-ranked descriptors, as shown in Table . This finding suggested that compounds with this descriptor are prone to be active NAIs against influenza A. Moreover, nArCOOH was the following largest descriptor usage in model construction, which tend to be absent in active NAIs. The propensity score of nArCOOH suggested that there are differences in functional group occurrence among active and inactive NAIs, in which the majority of this descriptor was found in inactive NAIs against neuraminidase of influenza A.

The decision tree model for classifying NAIs against influenza B exhibits high predictive performance, with 87.79% accuracy, 87.18% sensitivity, 88.04% specificity and 0.72 MCC. The descriptors used in model construction consisted of the number of aliphatic tertiary sp<sup>2</sup> carbons (nR=Ct), number of secondary alcohols (nOHs), number of aliphatic esters (nRCOOR), number of aliphatic ethers (nROR), number



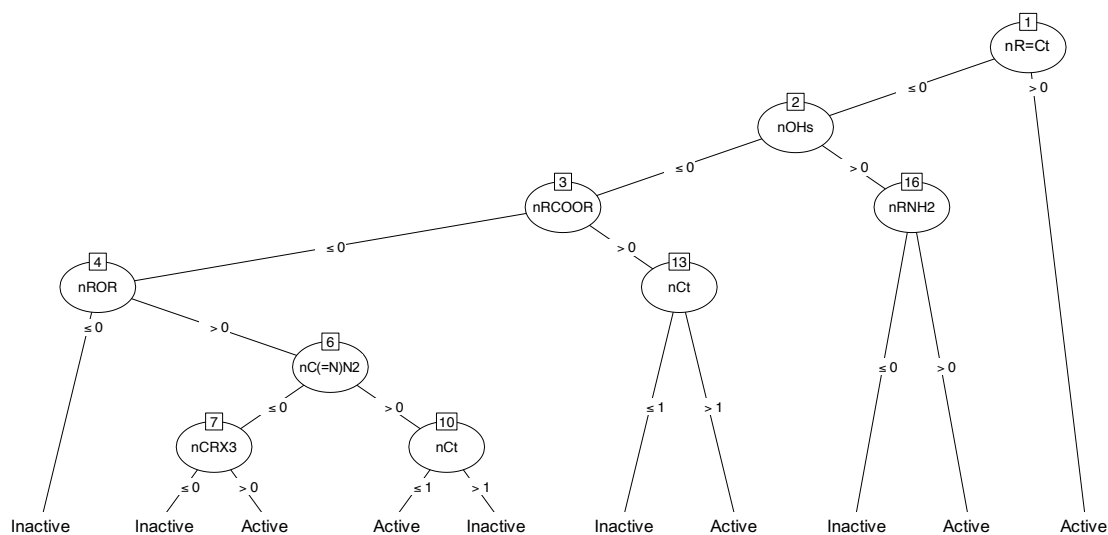
**Figure 6** Illustration of decision tree model for classifying the activity of NAIs against influenza type A according to their functional group descriptors.

of guanidine derivatives ( $nC(=N)N_2$ ), number of  $CRX_3$  ( $nCRX_3$ ), number of total tertiary  $sp^3$  carbons ( $nCt$ ) and number of aliphatic primary amines ( $nRNH_2$ ), with percent of descriptor usage of 74.81%, 70.23%, 66.41%, 36.64%, 30.53%, 9.92% and 4.58%, respectively. The largest usage of functional group descriptors was observed as  $nR=Ct$ , which was the root node of the decision tree model, and compounds that possessed  $nR=Ct > 0$  were prone to be active NAIs according to the model summarized in Figure 7. Interestingly,  $nR=Ct$  was located in the top-ranked propensity score, as shown in Table , which supported the previous assumption. Furthermore, compounds lacking  $nCRX_3$  tend to be classified as inactive compounds. The propensity score of these descriptors also suggested that the majority of this descriptor was found in active NAIs. These results indicate the importance of functional groups inside NAIs molecules that facilitate bioactivity against influenza B neuraminidase. Using a combination of decision tree and propensity score can provide insights regarding the important functional groups relevant to the bioactivity of NAIs.

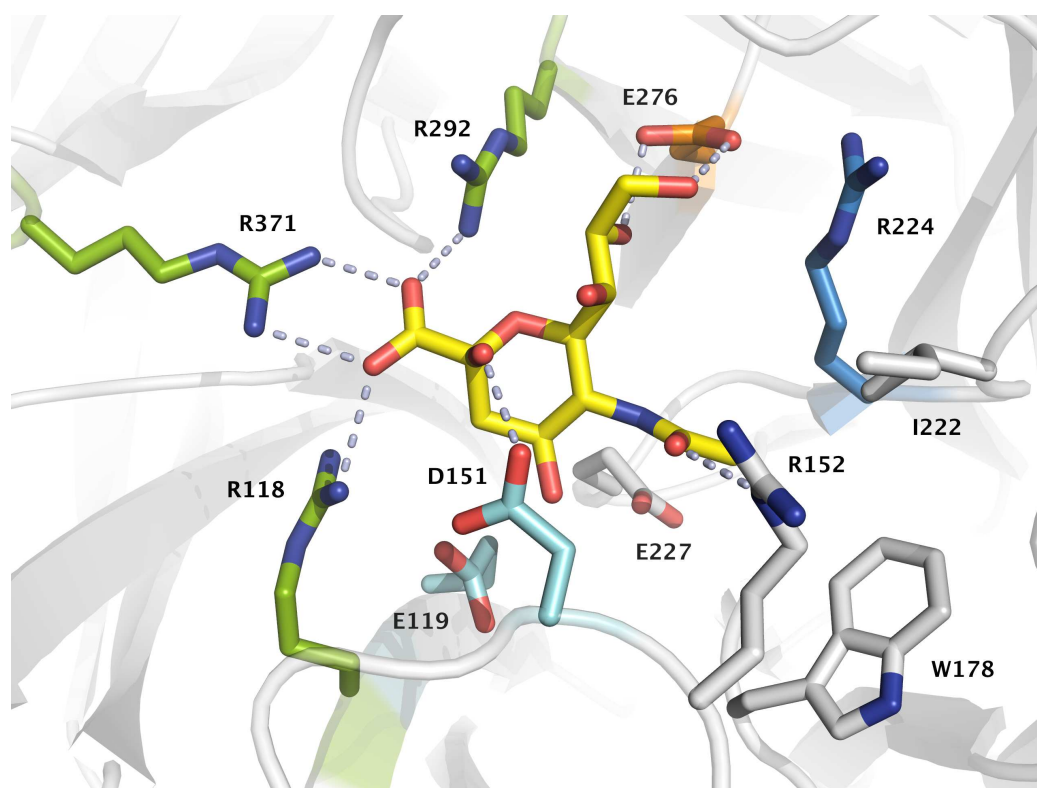
## Binding mode analysis

The observations on the active set of NAIs fragments revealed a pattern of molecular scaffolds that exhibited activity against the NA glycoprotein of type A influenza. Note that the molecules shared a similar conformation substructure as the original substrate, sialic acid, and tended to exhibit inhibitory potential against this enzyme. The binding pocket in the active site of NA contains eight highly conserved amino acid residues, which interact with the substrate and provide catalytic activity in the binding pocket. These residues can be grouped into five minor sites, as illustrated in Figure 8. Thus, designing novel NAIs requires choosing functional groups that can interact and fit with these sites of conserved residues to prevent catalytic reactions with this enzyme. To investigate the binding modes of active compounds against neuraminidase of both influenza type A and B, a combination of molecular docking and post-docking analysis using AutoDock Vina (Trott and Olson, 2010) and SiMMAP web-server (Chen et al., 2010), respectively, was employed to identify key interactions and important moieties facilitating protein-ligand interactions.

The analysis of 148 active NAIs against influenza A revealed four distinct binding anchors (Elec1, vdW1, vdW2 and vdW3) with their site-moiety preferences. Elec1 is the first anchor site, and it facilitates electrostatic interactions with carboxylic and alkyl phosphate groups of NAIs through the positive charge of the arginine side chain. The amino acid members of this anchor include R118, R292 and R371, which are responsible for the S1 subsite of influenza A neuraminidase active site (von Itzstein, 2011; Stoll et al., 2003). In contrast, another three anchor sites are facilitated by van der Waals interactions. The first anchor, vdW1, consisted of R152, I222 and E227, which are responsible for the S3 subsite of the neuraminidase binding pocket (von Itzstein, 2011; Stoll et al., 2003). The moiety preferences of this



**Figure 7** Illustration of decision tree model for classifying the activity of NAIs against influenza type B according to their functional group descriptors.



**Figure 8** The interactions of  $\alpha$ -Neu5Ac with conserved amino acid residues inside the binding pocket of influenza A/N2 neuraminidase (PDB ID: 2BAT). Amino acid residues in subsites S1, S2, S3, S4 and S5 are labeled by green, cyan, white, blue and orange, whereas the hydrogen bonds connecting ligand moieties and amino acid side chains are indicated by dashed lines.

anchor are composed of a heterocyclic ring, aromatic moiety, phenol group and aliphatic moiety with alkene. Another van der Waals interaction site was found at the vdW2 anchor site, which contains R224, E227 and R292 as key residues. This anchor facilitates van der Waals contact against aliphatic moieties with an alkene linkage, heterocyclic and aromatic moieties. vdW3 is the final anchor site, with a moiety preference of heterocyclic moiety, alkene linkage of aliphatic moiety and formamidine group. These findings have shown that NAIs interact with both functional residues that facilitate enzymatic reactions and structural residues that maintain the active site architecture (Shtyrya et al., 2009).

The analysis of the binding anchor of 45 active NAIs targeting influenza type B revealed four different anchor sites of the binding pocket: Elec1, Hbond1, vdW1 and vdW2. Electrostatic interactions between NAIs and amino acid residues primarily occurred with R115, R291 and R373 (comparable to R118, R292 and R371 in N2 numbering), which are members of the Elec1 anchor. The positive charge of the arginine side chain prefers carboxylic groups as its moiety preference. Note that this finding is similar to anchor Elec1 of influenza A neuraminidase. Interestingly, there are several moiety types of NAIs against influenza B virus that tend to form hydrogen bonds with amino acids in the Hbond1 anchor. The phenolic moiety of D148 and the carboxylic side chain of Y408 (comparable to D151 and Y406 in N2 numbering) facilitate hydrogen bonding through amino groups, carboxylic moieties, primary and secondary alcohols and ester moieties. It can be observed that these residues are members of the S2 subsite of influenza neuraminidase and are responsible for catalytic residues essential for enzyme functioning (Shtyrya et al., 2009). Furthermore, van der Waals contact sites are observed at two anchor site: vdW1 and vdW2. The first van der Waals interaction site is facilitated by R149, W176 and R222 (comparable to R152, W178 and R224 in N2 numbering), and their moiety preference is aliphatic moiety with alkene linkage, heterocyclic ring and aromatic moiety. The second van der Waals anchor is facilitated by I220, R222 and E274 (comparable to I222, R224 and E276 in N2 numbering), which have a heterocyclic ring and alkene linkage of aliphatic moiety as their moiety preference. The results of the post-docking analysis revealed the important amino acid residues and their moiety preferences that can generate potential protein-inhibitor complexes to inhibit enzymatic functioning of influenza neuraminidase.



**Table 11** Propensity score and percentage of functional group compositions in active and inactive neuraminidase inhibitors against influenza type A

Functional group	Description	Percentage of Frequency (%)			Propensity score (Rank)
		Active	Inactive	Active - Inactive	
nCs	Number of total secondary C(sp <sup>3</sup> )	22.59	9.27	13.32	1000 (1)
nCrS	Number of ring secondary C(sp <sup>3</sup> )	15.20	5.45	9.74	878 (2)
nCp	Number of terminal primary C(sp <sup>3</sup> )	8.24	5.01	3.23	655 (3)
nRCONHR	Number of secondary amides (aliphatic)	3.84	1.54	2.30	624 (4)
nRCOOH	Number of carboxylic acids (aliphatic)	3.54	1.32	2.22	621 (5)
nRNH <sub>2</sub>	Number of primary amines (aliphatic)	3.64	1.74	1.90	610 (6)
nROR	Number of ethers (aliphatic)	2.35	1.07	1.27	589 (7)
nRCONR <sub>2</sub>	Number of tertiary amides (aliphatic)	1.62	0.37	1.25	588 (8)
nCconj	Number of non-aromatic conjugated C(sp <sup>2</sup> )	8.63	7.50	1.14	584 (9)
nCt	Number of total tertiary C(sp <sup>3</sup> )	1.92	0.90	1.02	580 (10)
nCrt	Number of ring tertiary C(sp <sup>3</sup> )	1.65	0.62	1.02	580 (11)
nRNHR	Number of secondary amines (aliphatic)	1.17	0.35	0.82	573 (12)
nR=Ct	Number of aliphatic tertiary C(sp <sup>2</sup> )	1.47	0.77	0.70	569 (13)
nC(=N)N <sub>2</sub>	Number of guanidine derivatives	1.20	0.70	0.50	562 (14)
nP(=O)O <sub>2</sub> R	Number of phosphonates (thio-)	0.10	0.00	0.10	548 (15)
nCRX <sub>3</sub>	Number of CRX <sub>3</sub>	0.10	0.00	0.10	548 (16)
nN=C-N<	Number of amidine derivatives	0.05	0.00	0.05	547 (17)
nRSR	Number of sulfides	0.05	0.00	0.05	547 (18)
nArX	Number of X on aromatic ring	0.05	0.00	0.05	547 (19)
nR=Cs	Number of aliphatic secondary C(sp <sup>2</sup> )	4.44	4.41	0.03	546 (20)
nR=Cp	Number of terminal primary C(sp <sup>2</sup> )	0.12	0.10	0.03	546 (21)
nOxolanes	Number of Oxolanes	0.05	0.02	0.03	546 (22)
nCq	Number of total quaternary C(sp <sup>3</sup> )	0.02	0.00	0.02	546 (23)
nCrq	Number of ring quaternary C(sp <sup>3</sup> )	0.02	0.00	0.02	546 (24)
nC=N-N<	Number of hydrazones	0.02	0.00	0.02	546 (25)
nCXr=	Number of X on ring C(sp <sup>2</sup> )	0.02	0.00	0.02	546 (26)
nCconjX	Number of X on exo-conjugated C	0.02	0.00	0.02	546 (27)
nAzetidines	Number of Azetidines	0.02	0.00	0.02	546 (28)
nBeta-Lactams	Number of Beta-Lactams	0.02	0.00	0.02	546 (29)
nImidazoles	Number of Imidazoles	0.02	0.00	0.02	546 (30)
nTriazoles	Number of Triazoles	0.05	0.05	0.00	545 (31)

**Table 11** Continued ...

Functional group	Description	Percentage of Frequency (%)			Propensity score (Rank)
		Active	Inactive	Active - Inactive	
nN+	Number of positively charged N	0.02	0.02	0.00	545 (32)
nROH	Number of hydroxyl groups	4.12	4.13	-0.02	544 (33)
nRCONH <sub>2</sub>	Number of primary amides (aliphatic)	0.00	0.02	-0.02	544 (34)
nArCONHR	Number of secondary amides (aromatic)	0.00	0.02	-0.02	544 (35)
nRNO <sub>2</sub>	Number of nitro groups (aliphatic)	0.00	0.02	-0.02	544 (36)
nSO	Number of sulfoxides	0.00	0.02	-0.02	544 (37)
nS(=O) <sub>2</sub>	Number of sulfones	0.00	0.02	-0.02	544 (38)
nROCON	Number of (thio-) carbamates (aliphatic)	0.02	0.07	-0.05	543 (39)
nOHt	Number of tertiary alcohols	0.00	0.07	-0.07	542 (40)
nFuranes	Number of Furanes	0.00	0.07	-0.07	542 (41)
nArCNO	Number of oximes (aromatic)	0.00	0.10	-0.10	542 (42)
nRCOOR	Number of esters (aliphatic)	0.10	0.22	-0.12	541 (43)
nArNH <sub>2</sub>	Number of primary amines (aromatic)	0.00	0.15	-0.15	540 (44)
nPyridines	Number of Pyridines	0.00	0.15	-0.15	540 (45)
nPyrrolidines	Number of Pyrrolidines	0.20	0.37	-0.17	539 (46)
nCONN	Number of urea (-thio) derivatives	0.02	0.22	-0.20	538 (47)
nSO <sub>2</sub> N	Number of sulfonamides (thio-/dithio-)	0.00	0.22	-0.22	537 (48)
nArNHR	Number of secondary amines (aromatic)	0.00	0.45	-0.45	530 (49)
nOHp	Number of primary alcohols	0.17	0.70	-0.52	527 (50)
nRNR <sub>2</sub>	Number of tertiary amines (aliphatic)	0.50	1.17	-0.67	522 (51)
nOHs	Number of secondary alcohols	0.25	1.00	-0.75	519 (52)
nArCOOH	Number of carboxylic acids (aromatic)	0.00	0.87	-0.87	515 (53)
nArOR	Number of ethers (aromatic)	0.10	1.02	-0.92	513 (54)
nArCO	Number of ketones (aromatic)	0.00	1.32	-1.32	500 (55)
nArOH	Number of aromatic hydroxyls	0.05	3.44	-3.39	429 (56)
nCbH	Number of unsubstituted benzene C(sp <sup>2</sup> )	4.69	10.34	-5.64	352 (57)
nCb-	Number of substituted benzene C(sp <sup>2</sup> )	1.30	10.44	-9.14	233 (58)
nCar	Number of aromatic C(sp <sup>2</sup> )	6.16	22.12	-15.95	0 (59)
		<b>100</b>	<b>100</b>		<b>R = 1.00</b>

**Table 12** Propensity score and percentage of functional group compositions in active and inactive neuraminidase inhibitors against influenza type A

Functional group	Description	Percentage of Frequency (%)			Propensity score (Rank)
		Active	Inactive	Active - Inactive	
nCrS	Number of ring secondary C(sp <sup>3</sup> )	18.14	11.04	7.11	1000 (1)
nCs	Number of total secondary C(sp <sup>3</sup> )	24.25	17.98	6.27	952 (2)
nCconj	Number of non-aromatic conjugated C(sp <sup>2</sup> )	10.83	6.35	4.48	848 (3)
nR=Ct	Number of aliphatic tertiary C(sp <sup>2</sup> )	2.67	0.08	2.58	738 (4)
nRNR <sub>2</sub>	Number of tertiary amines (aliphatic)	1.55	0.13	1.42	670 (5)
nR=Cs	Number of aliphatic secondary C(sp <sup>2</sup> )	4.64	3.64	1.01	646 (6)
nRCOOH	Number of carboxylic acids (aliphatic)	3.87	2.93	0.94	642 (7)
nROH	Number of hydroxyl groups	4.90	4.06	0.85	637 (8)
nCp	Number of terminal primary C(sp <sup>3</sup> )	8.43	7.69	0.73	630 (9)
nRNH <sub>2</sub>	Number of primary amines (aliphatic)	3.87	3.14	0.73	630 (10)
nOHs	Number of secondary alcohols	0.60	0.17	0.43	613 (11)
nCRX <sub>3</sub>	Number of CRX <sub>3</sub>	0.17	0.00	0.17	598 (12)
nAzetidines	Number of Azetidines	0.17	0.00	0.17	598 (13)
nOHp	Number of primary alcohols	0.43	0.33	0.10	593 (14)
nPyrrolidines	Number of Pyrrolidines	0.26	0.17	0.09	593 (15)
nCq	Number of total quaternary C(sp <sup>3</sup> )	0.09	0.00	0.09	593 (16)
nCrq	Number of ring quaternary C(sp <sup>3</sup> )	0.09	0.00	0.09	593 (17)
nCXr=	Number of X on ring C(sp <sup>2</sup> )	0.09	0.00	0.09	593 (18)
nCconjX	Number of X on exo-conjugated C	0.09	0.00	0.09	593 (19)
nOxolanes	Number of Oxolanes	0.09	0.00	0.09	593 (20)
nROR	Number of ethers (aliphatic)	1.98	1.92	0.05	591 (21)
nRCOOR	Number of esters (aliphatic)	0.17	0.13	0.05	590 (22)
nR=Cp	Number of terminal primary C(sp <sup>2</sup> )	0.09	0.04	0.04	590 (23)
nRCONH <sub>2</sub>	Number of primary amides (aliphatic)	0.00	0.04	-0.04	585 (24)
nCONN	Number of urea (-thio) derivatives	0.00	0.04	-0.04	585 (25)
nN+	Number of positively charged N	0.00	0.04	-0.04	585 (26)
nArOH	Number of aromatic hydroxyls	0.00	0.04	-0.04	585 (27)
nRSR	Number of sulfides	0.00	0.04	-0.04	585 (28)
nSO <sub>2</sub> N	Number of sulfonamides (thio-/dithio-)	0.00	0.04	-0.04	585 (29)
nRCONHR	Number of secondary amides (aliphatic)	3.96	4.01	-0.06	584 (30)
nArNH <sub>2</sub>	Number of primary amines (aromatic)	0.00	0.08	-0.08	583 (31)

**Table 12** Continued ...

Functional group	Description	Percentage of Frequency (%)			Propensity score (Rank)
		Active	Inactive	Active - Inactive	
nC=N-N<	Number of hydrazones	0.00	0.08	-0.08	583 (32)
nArX	Number of X on aromatic ring	0.00	0.08	-0.08	583 (33)
nROCON	Number of (thio-) carbamates (aliphatic)	0.00	0.13	-0.13	580 (34)
nTriazoles	Number of Triazoles	0.00	0.17	-0.17	578 (35)
nArOR	Number of ethers (aromatic)	0.00	0.21	-0.21	576 (36)
nArNHR	Number of secondary amines (aromatic)	0.00	0.25	-0.25	573 (37)
nRNHR	Number of secondary amines (aliphatic)	0.69	1.13	-0.44	562 (38)
nArCOOH	Number of carboxylic acids (aromatic)	0.00	0.59	-0.59	554 (39)
nC(=N)N <sub>2</sub>	Number of guanidine derivatives	0.60	1.34	-0.74	545 (40)
nRCONR <sub>2</sub>	Number of tertiary amides (aliphatic)	0.69	2.01	-1.32	511 (41)
nCrt	Number of ring tertiary C(sp <sup>3</sup> )	0.60	2.38	-1.78	484 (42)
nCt	Number of total tertiary C(sp <sup>3</sup> )	0.86	2.76	-1.90	478 (43)
nCb-	Number of substituted benzene C(sp <sup>2</sup> )	0.43	3.39	-2.96	416 (44)
nCbH	Number of unsubstituted benzene C(sp <sup>2</sup> )	2.15	8.65	-6.50	210 (45)
nCar	Number of aromatic C(sp <sup>2</sup> )	2.58	12.71	-10.13	0 (46)
		<b>100</b>	<b>100</b>		<b>R = 1.00</b>

## CONCLUSION

The emergence of novel influenza strains that possess resistance mutations emphasize the importance of finding novel therapeutic agents for treatment and prophylaxis. The increase in the emergence of influenza viruses, particularly mutant variants, calls for the development of novel promising NAIs, in addition to the three currently approved NAIs, for preparedness against influenza. Nevertheless, there are several compounds that were tested to evaluate their inhibitory activity against influenza neuraminidase. Expanding the chemical space available in public databases of NAIs provides an opportunity to investigate the molecular factors relevant to the bioactivity of NAIs. In addition, a combination of various computational approaches revealed the structure-activity relationships of NAIs, which are essential for rational drug design to develop new promising therapeutic agents against influenza neuraminidase. Therefore, this work reports a large-scale study of the chemical space of NAIs against influenza type A and B and performs statistical and QSAR investigations of both molecular and quantum chemical properties that contribute inhibitory activity against influenza neuraminidase. Moreover, maximum common molecular substructures and their functional groups were analyzed from a ligand-based perspective. In addition, the binding modes of active NAIs were investigated to observe important amino acid residues and their site-moiety preferences that facilitate protein-ligand interaction. Moreover, informative descriptors leading to good performance of the QSAR model were achieved in combination with a statistical analysis that revealed the molecular properties that distinguish between active and inactive classes of NAIs. The molecular properties of the active group include a higher number of rotatable bonds, number of hydrogen-bond donors and acceptor atoms, total energy of molecules and kinetic stability. In addition, the active group also appeared to possess fewer cyclic rings and lower lipophilicity and charge according to the univariate analysis results. The maximum common substructures observed in NAIs are primarily cyclohexene-based, dihydropyran-based and cyclopentane-based scaffolds in the molecular framework. These fragments were suggested to be the privileged structures that contribute to neuraminidase inhibition. Functional group analysis revealed the important functional groups and their characteristic patterns among active and inactive compounds. The results of decision tree models suggested that the bioactivity of NAIs can be classified according to their functional groups, which highlights the importance of functional groups incorporated inside molecules. Furthermore, the results of the binding mode analysis revealed key interactions that facilitated protein-ligand binding and their moiety preferences. Thus, these finding may provide insights regarding important molecular properties and essential molecular structures for the development of novel neuraminidase inhibitors.

## ACKNOWLEDGMENTS

This research project is supported by the Office of Higher Education Commission and Mahidol University under the National Research Universities Initiative.

## REFERENCES

- Babu, Y. S., Chand, P., Bantia, S., Kotian, P., Dehghani, A., El-Kattan, Y., Lin, T. H., Hutchison, T. L., Elliott, A. J., Parker, C. D., Ananth, S. L., Horn, L. L., Laver, G. W., and Montgomery, J. A. (2000). Bcx-1812 (rwj-270201): discovery of a novel, highly potent, orally active, and selective influenza neuraminidase inhibitor through structure-based drug design. *Journal of Medicinal Chemistry*, 43(19):3482–6.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–42.
- Charoenkwan, P., Shoombuatong, W., Lee, H.-C., Chaijaruwanich, J., Huang, H.-L., and Ho, S.-Y. (2013). Scmcrys: Predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of p-collocated amino acid pairs. *PLoS ONE*, 8(9):e72368.
- ChemAxon Ltd. (2015a). LibMCS, version 15.1.12.0.
- ChemAxon Ltd. (2015b). MolConverter, version 15.1.12.0.
- Chen, Y.-F., Hsu, K.-C., Lin, S.-R., Wang, W.-C., Huang, Y.-C., and Yang, J.-M. (2010). SiMMap: a web server for inferring site-moiety map to recognize interaction preferences between protein pockets and compound moieties. *Nucleic Acids Research*, 38(suppl 2):W424–W430.
- Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P.,

- Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery Jr, J. A., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, J. M., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, Ö., Foresman, J. B., Ortiz, J. V., Cioslowski, J., and Fox, D. J. (2009). Gaussian'09 Revision A.01. Gaussian Inc. Wallingford CT 2009.
- Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, 130(12):995–1004.
- Greene, N. and Naven, R. (2009). Early toxicity screening strategies. *Current Opinion in Drug Discovery & Development*, 12(1):90–7.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.
- Itzstein, M. and Thomson, R. (2009). Anti-influenza drugs: The development of sialidase inhibitors. In Kräusslich, H.-G. and Bartenschlager, R., editors, *Antiviral Strategies*, volume 189 of *Handbook of Experimental Pharmacology*, pages 111–154. Springer Berlin Heidelberg.
- Jolliffe, I. (2005). *Principal Component Analysis*. John Wiley & Sons, Ltd.
- Kim, C. U., Lew, W., Williams, M. A., Liu, H., Zhang, L., Swaminathan, S., Bischofberger, N., Chen, M. S., Mendel, D. B., Tai, C. Y., Laver, W. G., and Stevens, R. C. (1997). Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site: design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity. *Journal of the American Chemical Society*, 119(4):681–90.
- Kuhn, M., Weston, S., Coulter, N., and Culp, M. (2015). C50: C5.0 decision trees and rule-based models.
- Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1).
- Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46(1-3):3–26.
- Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 35(Database issue):D198–201.
- Meindl, P., Bodo, G., Palese, P., Schulman, J., and Tuppy, H. (1974). Inhibition of neuraminidase activity by derivatives of 2-deoxy-2,3-dehydro-n-acetylneuraminic acid. *Virology*, 58(2):457–63.
- Nantasenamat, C., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2010). Advances in computational methods to predict the biological activity of compounds. *Expert Opinion on Drug Discovery*, 5(7):633–54.
- Nantasenamat, C., Li, H., Mandi, P., Worachartcheewan, A., Monnor, T., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2013). Exploring the chemical space of aromatase inhibitors. *Molecular Diversity*, 17(4):661–77.
- O’Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33.
- Peasah, S. K., Azziz-Baumgartner, E., Breese, J., Meltzer, M. I., and Widdowson, M. A. (2013). Influenza cost and cost-effectiveness studies globally—a review. *Vaccine*, 31(46):5339–48.
- Prachayasittikul, V., Worachartcheewan, A., Shoombuatong, W., Songtawee, N., Simeon, S., Prachayasittikul, V., and Nantasenamat, C. (2015). Computer-aided drug design of bioactive natural products. *Current Topics in Medicinal Chemistry*, 15(18):1780–800.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Shoombuatong, W., Prachayasittikul, V., Anuwongcharoen, N., Songtawee, N., Monnor, T., Prachayasittikul, S., Prachayasittikul, V., and Nantasenamat, C. (2015a). Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. *Journal of Drug Design, Development and Therapy*, 9:4515–49.
- Shoombuatong, W., Prachayasittikul, V., Prachayasittikul, V., and Nantasenamat, C. (2015b). Prediction of aromatase inhibitory activity using the efficient linear method (ELM). *EXCLI Journal*, 14:452–464.



- 736 Shtyrya, Y. A., Mochalova, L. V., and Bovin, N. V. (2009). Influenza virus neuraminidase: structure and  
737 function. *Acta Naturae*, 1(2):26–32.
- 738 Stevens, A. (2014). An introduction to the prospector package. pages 1–22.
- 739 Stoll, V., Stewart, K. D., Maring, C. J., Muchmore, S., Giranda, V., Gu, Y. G., Wang, G., Chen, Y., Sun,  
740 M., Zhao, C., Kennedy, A. L., Madigan, D. L., Xu, Y., Saldivar, A., Kati, W., Laver, G., Sowin, T.,  
741 Sham, H. L., Greer, J., and Kempf, D. (2003). Influenza neuraminidase inhibitors: structure-based  
742 design of a novel inhibitor series. *Biochemistry*, 42(3):718–27.
- 743 Talete srl. (2007). Dragon for Windows (Software for Molecular Descriptor Calculations), version 5.5.
- 744 Taylor, N. R. and von Itzstein, M. (1994). Molecular modeling studies on ligand binding to sialidase from  
745 influenza virus and the mechanism of catalysis. *Journal of Medicinal Chemistry*, 37(5):616–24.
- 746 Trott, O. and Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a  
747 new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*,  
748 31(2):455–61.
- 749 Tuna, N., Karabay, O., and Yahyaoglu, M. (2012). Comparison of efficacy and safety of oseltamivir and  
750 zanamivir in pandemic influenza treatment. *Indian Journal of Pharmacology*, 44(6):780–3.
- 751 Varghese, J. N., McKimm-Breschkin, J. L., Caldwell, J. B., Kortt, A. A., and Colman, P. M. (1992).  
752 The structure of the complex between influenza virus neuraminidase and sialic acid, the viral receptor.  
753 *Proteins: Structure, Function, and Bioinformatics*, 14(3):327–332.
- 754 von Itzstein, M. (2011). *Influenza Virus Sialidase-A Drug Discovery Target*. Springer Science & Business  
755 Media.
- 756 von Itzstein, M., Dyason, J. C., Oliver, S. W., White, H. F., Wu, W. Y., Kok, G. B., and Pegg, M. S.  
757 (1996). A study of the active site of influenza virus sialidase: an approach to the rational design of  
758 novel anti-influenza drugs. *Journal of Medicinal Chemistry*, 39(2):388–91.
- 759 von Itzstein, M., Wu, W. Y., Kok, G. B., Pegg, M. S., Dyason, J. C., Jin, B., Van Phan, T., Smythe, M. L.,  
760 White, H. F., Oliver, S. W., and et al. (1993). Rational design of potent sialidase-based inhibitors of  
761 influenza virus replication. *Nature*, 363(6428):418–23.
- 762 Worachartcheewan, A., Nantasenamat, C., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2013).  
763 Quantitative population-health relationship (QPHR) for assessing metabolic syndrome. *EXCLI Journal*,  
764 12:569–583.
- 765 World Health Organization (2014). Influenza (Seasonal). Available at  
766 <http://www.who.int/mediacentre/factsheets/fs211/en/> (accessed 19 May 2015).
- 767 Yamamoto, T., Kumazawa, H., Inami, K., Teshima, T., and Shiba, T. (1992). Syntheses of sialic acid  
768 isomers with inhibitory activity against neuraminidase. *Tetrahedron Letters*, 33(39):5791 – 5794. The  
769 International Journal for the Rapid Publication of Preliminary Communications in Organic Chemistry.