QSAR study of N-substituted Oseltamivir derivatives as potent avian influenza virus H5N1 inhibitors using quantum chemical descriptors and statistical methods

Samir Chitita*<sup>a,b</sup>, Adnane Aouidate<sup>c</sup>, Assia Belhassan<sup>b</sup>, Abdellah Oussaa<sup>b</sup>, Abdelali Idrissi Taourati<sup>b</sup>, Bouhya Eldrissi<sup>p</sup>, Mounir Ghamali<sup>b</sup>, Mohammed Bouachrine<sup>b</sup>, Tahar Lakhlifi<sup>b</sup>

Supporting information

**Note 1:** The coefficient of determination ($R^2$), which is the proportion of variance (%) in the dependent variable that can be explained by the independent variable [1].

$$R^2 = 1 - \frac{\sum (Y_{obs} - Y_{calc})^2}{\sum (Y_{obs} - \bar{Y}_{obs})^2}$$

$Y_{obs}$: The observed response value.
$Y_{calc}$: The model derived calculated/predicted response.
$\bar{Y}_{obs}$: The average of the observed response values.

**Note 2:** The adjusted coefficient of determination ($R_{adj}^2$) used to reflect the explained variance (the fraction of the data variance explained by the model) in a better way [1].

$$R_{adj}^2 = \frac{R^2(N - 1) - p}{N - p - 1}$$

$p$ is the number of predictor variables used in the model development and $N$ is the number of compounds.

**Note 3:** The means of the squares of the errors (or residuals) of the model (MSE or MSR): measure the average squared difference between the predicted and experimental activities values. The MSE is always strictly positive, and a good model will be with MSE values closer to zero [2].

$$MSE = \frac{1}{N}\sum(Y_{obs} - Y_{calc})^2$$

**Note 4:** The coefficient of Fischer (Fisher statistics $F$) used to judge the overall significance of the regression coefficients, the variance ratio. For overall significance of the regression coefficients, the $F$-value should be high [1].

$$F = \frac{\sum(Y_{calc} - \bar{Y}_{calc})^2}{N - p - 1}$$

$Y_{calc}$: The average of the predicted response values.

**Note 5:** The variance inflation factor VIF [3] to detect the absence of the multicollinearity between descriptors was used; models with descriptors correlated with each other are not significant. The VIF was defined as $1/(1 - R^2)$, where $R$ is the coefficient of correlation between one descriptor and all other descriptors in model. A VIF value greater than 5.0 indicates that the model is unstable; a value between 1.0 and 4.0 indicates that the model is acceptable [4].

The internal validations were performed by the Leave-One-Out cross validation method ($Q_{cv}^2$) and the Y-randomization test ($R_{Rand}^2$) and the external validation was evaluated by a test set ($R_{test}^2$).

**Note 6:** The internal validation procedure (leave-one-out cross validation) was employed, in which one compound is removed and the rebuilt model with the remaining molecules is used to predict the response of the eliminated compound. This one is then returned and a second is removed, and the cycle is repeated, and so on until all compounds have been removed one by one, and an overall correlation coefficient $Q_{cv}^2$ is computed [5]. A model is considered acceptable when the value of $Q_{cv}^2$ exceeds 0.6 [4, 6].
\[ Q_{cv}^2 = 1 - \frac{\sum (Y_{Calc} - Y_{Obs})^2}{\sum (Y_{Obs} - \bar{Y}_{Obs})^2} \]

- \( Y_{Obs} \): The observed response value.
- \( Y_{Calc} \): The model derived calculated response.
- \( Y_{Obs} \): The average of the observed response values.

**Note 7:** In order to assess the significance of the models and its accurate prediction ability for new compounds, an external prediction is necessary. This one remains the only way to determine both the generalizability of QSAR model for new chemicals and the true predictive power of the models. In this external validation, the obtained model was used to predict the activities of a test set comprising compounds that are similar to though not used in the training set and a correlation coefficient \( R^2_{test} \) is computed. This is usually performed by splitting a data set into a training set a test set, typically in a 1:5 ratio. For a predictive QSAR model, the value of \( R^2_{test} \) should be more than 0.5 [7, 8].

\[ R^2_{test} = 1 - \frac{\sum (Y_{Calc(test)} - Y_{Obs(test)})^2}{\sum (Y_{Obs(test)} - \bar{Y}_{Obs(training)})^2} \]

**Note 8:** Additional, the y-randomization approach was performed to ensure the robustness of a predictive model. Often, it is used along with the cross-validation. It consists of repeating the calculation procedure with randomized activities and subsequent probability assessment of the resultant statistics. The dependent variable vector is randomly shuffled and a new predictive model is developed using the original independent variable matrix. The new predictive models (after several repetitions) are expected to have low \( R^2 \) and \( Q_{cv}^2 \) values. If the opposite happens, then an acceptable model cannot be obtained for the specific modelling method and data [9]. Another parameter, \( cR_p^2 \) is also calculated which should be more than 0.5 for passing this test:

\[ cR_p^2 = R \times \sqrt{(R^2 - \text{Average } R_{Rand})^2} \] Where: Average \( R_{Rand} \) = average \( 'R' \) of random models [10].

---